Morphological Analysis of Gĩkũyũ using a Finite State Machine

[1]Kamau Chege [1]Wanjiku Ng'ang'a
[1]Peter W. Wagacha
[1]School of Computing and Informatics, University of Nairobi, Kenya.
kamauchege@gmail.com, wanjiku.nganga@uonbi.ac.ke, waiganjo@uonbi.ac.ke

[2]Guy De Pauw , [3]Jayne Mutiga
[2]CLiPS - Computational Linguistics Group, University of Antwerp, Belgium
[3]Centre for Translation and Interpretation, University of Nairobi, Kenya.
guy.depauw@ua.ac.be, jaynemutiga@yahoo.co.uk

Abstract

In this paper we present the development of a morphological analysis system for Gĩkũyũ Major morphological processes prevalent in Gĩkũyũ language are explored. These include verb inflection, verb and noun derivation and verb reduplication. In this work, finite state transducers are used to model Gĩkũyũ morphology. Xerox finite state tools are used to build the lexical transducers for the system and to model rewrite rules prevalent in Gĩkũyũ.

The system achieves an acceptable representation of Gĩkũyũ morphology. It can correctly analyze Gĩkũyũ words with an accuracy of 56%. As Gĩkũyũ is highly inflections, ambiguity is a big challenge both in morphological analysis and machine translation.

1 Introduction

Morphological analysis is an important first step in many natural language processing tasks such as parsing, machine translation, information retrieval, part-of-speech tagging among others. The field of natural language processing has progressed considerably over recent years. Despite this, many African languages have been left behind. This status quo has been influenced by two major reasons. Firstly, many African countries have adopted European international languages such as English and French among others as their official languages. This makes local languages unpopular and therefore not economically viable to invest in. Secondly, many African languages are resource-scarce. There is very little, if any, digitized linguistic resources for these languages. Moreover, lack of financial resources and political will hinders creation of such linguistic resources from scratch. Gĩkũyũ is one of the many resource-scarce African languages. Rapid revolution in Information and Communication Technology is affecting the way we learn, socialize, and do business among other things. There is need to position our local languages in this new paradigm else they will be lost. This can only be achieved through availing digital resources for them; and linguistic tools are vital in creation and dissemination of such resources. Gĩkũyũ is a highly agglutinative Bantu language spoken by between 6-7 million speakers in Central Kenya.

2 Literature Review

2.1 Previous work

A number of research initiatives have gone into morphological analysis of African languages. Various approaches have been taken on morphological analysis. De Pauw and Wagacha (2007) use unsupervised methods to learn Gĩkũyũ morphology. In their work, maximum entropy learning is used to carry out automatic induction of shallow morphological features for Gĩkũyũ. This approach is desirable given that it achieves morphology with minimal human effort, but availability of sufficient corpus to facilitate learning is a challenge.

Other NLP research works on Gĩkũyũ language include automatic diacritic correction using grapheme-based memory model (De Pauw et.al, 2007) and a Gĩkũyũ text-to-speech system using Festival tool. On Swahili, a Bantu language closely related to Gĩkũyũ, an attempt at morphological analysis based on two-level morphology (Koskeniemmi, 1983) is carried out in a project named SWATWOL (Hurskainen, 2004) at the University of Helsinki, Finland. In this work,

Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

Xerox's TWOLC tool is used to implement two level rules.

Other research works at morphological analysis of African languages include Kĩkamba Morphological analyzer using rules, SwaMorph analyzer for Swahili using finite state technology among other efforts on Central and Southern African languages.

2.2 Gĩkũyũ Morphology

Gĩkũyũ is a language spoken by a Kenyan community of Kamba-Kikuyu subgroup of the Bantu origin, with approximately 7 million speakers living in Central Kenya. The language has six dialects and is lexically similar to closely related languages such as Chuka, Embu, Meru, Kamba.

Gĩkũyũ is a highly inflectional language with a complex word structure and phonemics. Like many other Bantu languages, Gĩkũyũ has 15 noun classes and two additional locative classes. The language also has a concord system formed around the noun classes. The language is also tonal, a major source of ambiguity.

Noun Morphology: Gĩkũyũ nouns can be grouped into two categories namely derived and underived nouns. Underived nouns consist of named entities. Derived nouns can be formed in two ways. Firstly, they are formed by affixing prefixes of diminutive, augmentative or collective to an underived noun. Examples are,

thitima -gĩ-thitima (a big light bulb)
mũndũ – ma-mũndũ (many big persons)
i-ndathi – rũ-ndathi (a bad/ugly) bun)

The second form of derived nouns is formed through nominalization of verbs. This involves circumfixing verb roots with a set of prefix and suffixes to give different meaning depending on the type of nominalization. Nominalization types include agentive (most prevalent), occasion, manner, locative e.t.c. Examples include,

mũ-thaak-i (player) i-ho-ero (Place of prayer)
i-geth-a (harvesting occasion) ga-thom-i (the small one who reads)

The membership of a noun to a noun class is determined slightly by its initial characters but is mainly determined by the concord system which it enforces on other parts of speech in a sentence. All Gĩkũyũ nouns, underived or otherwise, can be affixed with a suffix *-inĩ* with the effect of changing the meaning from a referential entity to a location.

Verb Morphology: Gĩkũyũ language is agglutinative. Dependent morphemes are affixed to the independent morpheme to derive a certain meaning to the surface verb. A typical Gĩkũyũ verb consists of a combination zero or more of the dependent morphemes and a mandatory dependent morpheme, with the final vowel also being mandatory. Figure 1 illustrates the verb morphology.

The simplest verb consists of the verb root and the final vowel. These are usually commands or directives. Subjunctive verb formations i.e. commands, can optionally take a plural marker *-i* or *-ni*. Examples of Verbs are shown below;

ma-thom-e (so that they read)
ci-ti-raa-tũ-meny-ag-a (they were not knowing us)
a-gaa-kenaken-ithi-ag-io (he/she will be made happy a little more)
nĩ-kĩ-mũ-hũr-ag-a (it usually beats him/her)
nd-aa-ngĩ-kaa-ma-caracar-ithi-ang-ia (he would not have help them a little more in searching)
kom-a-i (sleep)
rehe-ni (bring)

Reduplication: Gĩkũyũ verbs also undergo verb reduplication. This involves repeating part or the entire lexical root of the verb, depending on the number of syllables in the root. The meaning: *Focus+ Subj_Marker+ Neg+ Cond+ Tense+ Obj_Marker+ Redupl+ Verb+ Dev_Ext+ Aspect+ FV* derived from reduplication varies among verb stems but usually means repeatedly doing the action, doing the action for a little longer, among others.

Verbs with one or two syllables undergo full reduplication. Verbs with more than two syllables undergo partial reduplication. Only the first two syllables are repeated. In addition, the last vowel, whatever character it is, is rewritten as *'a'*. Examples are;

koma - koma-koma (sleep a little) ne – nea-nea (give a little)
negena – nega-negena (make noise a little more)
tiga – tiga-tiga (leave a little)
hungura – hunga-hungura (slip under a little more)

Verb Mutation: Gĩkũyũ verbs are also affected by consonantal and vowel phonemics.

Prenasalized stop formation (and its variant called Meinhof's Law) involves consonants *b, c, r, t, g, k* being replaced with NC composites in verbs obeying first person singular, noun classes 1, 8 and 9 concord systems. Vowel-initial verbs whose first consonant is any of the participating consonants also undergo this mutation. Examples include;

roota – ndoota ũma -nyũmia
guna – ng'una komia – ngomia
cuuka – njuuka egeka njegeka

Dahl's Law is a consonantal mutation that involves the voiceless trigger sound *–k* appearing before other voiceless sounds c,k,t being replaced with its equivalent voiced sound *–g-* . Examples include;
kũ-thoma – gũthoma ka-ka-thaaka – gagathaaka
ma-kaa-ka-menya - magaakamenya

Vowel mutation includes vowel lengthening before prenasalized stops and vowel assimilation when some vowel combinations appear in the neighborhood of each other.

2.3 Finite State Transducers

The ability to formally describe language lexicon and morphological rewrite rules using finite state machines have made it a popular approach to computational morphology. The two-level formalism has been successfully described using finite state transducers. Finite state networks model grammar rules which can be used to validate if a given input string belongs to a language.

Finite state transducers are used to model both morphological generators and analyzers. An FST generator uses rules to transform a lexical form into a surface form. An FST analyzer consults both rules and the lexicon to transform a surface string into the corresponding lexical form. See Figure 2.

Since FSTs are bidirectional, the two processes are usually the opposite of each other. Finite state transducers can also be used to carry out other NLP tasks such as tokenization, part-of-speech tagging among others.

2.4 Two-Level formalism

Transformation of input strings from lexical to surface representation involves a series of intermediate stages. Transitions between these stages involve zero or more generative rules called rewrite rules. Classical generative phonology implements rewrite rules in sequence. Koskeniemmi (1983) two-level formalism allows the two representations to be directly related with no need for intermediate steps. The formalism models the relation between the two formalism using parallel rules. The rules act a set of conditions, not actions, whose role is to ascertain whether the correspondence between the two representations is correct.

Lexical: NC g u i Phonemic: n g u i	Two-Level Rule: NC => n;

Figure 3: Two Level Representation

This formalism is bidirectional, which allows generation and analysis to be modeled in parallel. It can also be easily represented using finite state transducers.

2.5 Challenges

Gĩkũyũ language is more spoken than written (and read). This means that very few written sources exist for the language. Furthermore, the language use is losing favor to the more popular national languages; Swahili and English.

The Gĩkũyũ orthography includes two diacritically marked characters ĩ and ũ that represent different phonemes from i and u respectively. Standard keyboard lacks these two characters and hence many users tend to use their unmarked equivalents. In addition, the characters complicate automated corpus collection through such methods as OCR.
Gĩkũyũ language is also tonal; this is usually a major source of word ambiguity.

3 Methodology

3.1 Corpus Collection

This work uses a 25,000 word corpus, corrected from a variety of resources. Out of this, a set 19,000 words is from previous works (De Pauw and Wagacha, 2007; De Pauw et.al, 2007) carried out on Gĩkũyũ language. The set has a bias religious material but also includes text from hymn books, poems, short stories, novels and

also web crawling. A small set is also manually transcribed. The remaining set of 6,000 words is collected during this project and includes materials from manual transcription, constitutional review, online religious material, agricultural material, blog entries among others. It is mainly running text and part of it is held out as test data.

The data is cleaned and annotated manually and also automated by writing perl scripts. This involved removing non-Gĩkũyũ words, part-of-speech tagging and also grouping using structure similarity.

3.1 Xerox Tools

Xerox's Finite state tools, XFST and LEXC, were used to implement the Gĩkũyũ morphology system. Due to an encoding problem appearing during compile-replace for reduplication, the two diacritically marked characters are represented by two characters not in Gĩkũyũ alphabet. S represents ũ while L represents ĩ. The Gĩkũyũ lexicon files were developed using Xerox's lexicon format so as to be compiled used the LEXC compiler. The files were modularized along the major parts of speech.

3.2 Morphology System

The morphology system is modeled around Xerox's LEXC tool for creating large lexicon. Morphological markers are used to guide the meaning of each morpheme. Reduplication and rewrite rules are modeled using XFST tool and then composed with the lexicon after it has been compiled. The lexicon is implemented as continuation classes and organized around various parts of speech. Underived nouns, together with their diminutives, augmentatives, collectives and the locative modifier are implemented as a single lexicon as shown in figure 4.

Derived nouns are implemented in a separate lexicon file. To enforce circumfixation (prefix-suffix combination) associated with nominalization, flag diacritics are used, Gĩkũyũ verbs have a number of optional affixes. These are implemented through allowing a -ı- transition in every continuation class that is optional. To enhance the organization of the verb lexicon file, verbs roots are placed in different continuation classes depending on how structurally similar they are, as shown below.

LEXICON SPList (Have irregular endings)
he Suffs;
ne Suffs;

LEXICON IAMidLowList (ends with –ia, obeys mid-low sound)
endia:end Suffs;
rekia:rek Suffs;

LEXICON IAMidHighList (ends with –ia, obeys mid-high sound)
akia:ak Suffs;
giria:gir Suffs;

LEXICON ACausMidLowList (ends with –a and are causative, obeys mid-low sound)
cera:cer Suffs;
gema:gem Suffs;

LEXICON ACausMidHighList (ends with –a and are causative, obeys mid-high sound)
baca:bac Suffs;
gwata:gwat Suffs;

LEXICON AMidLowList (ends with –a, are not causative, obeys mid-low sound)
kombora:kombor Suffs;
hehenja:hehenj Suffs;

LEXICON AMidHighList (ends with –a, are not causative, obeys mid-high sound)
aka:ak Suffs;
kanya:kany Suffs;

Reduplication is implemented using compile-replace function in XFST tool. This is a 2-stage process. The first stage involves using regular expressions to identify sections of the verb to be reduplicated and replacing them with regular expressions as shown,

1. define *Redupl1 [[Syllables]^2] @-> "^[{ ... Z } ^2^] Z" ||"+[Redup]"[EFLAGS]* _[[Alpha]+ "+[Verb]"];*
2. define *Redupl2 [Syllables CON^{0,1}] @-> "^[{ ... Q } ^2 ^] Q" ||"+[Redup]"[EFLAGS]* _["+[Verb]"];*
3. define *Redupl3 [CON] @-> "^[{ ... V } ^2 ^] V" ||"+[Redup]"[EFLAGS]* _ ["+[Verb]"];*
4. define *Reduplication Redupl1 .o. Redupl2 .o. Redupl3;*

The compile-replace command is then invoked.

Rewrite rules for several aspects are also implemented using XFST. These include Dahl's Law, prenasalization, vowel assimilation, prenasals removal by diminutives, augmentatives and collectives among others.

1. define *DahlsLaw [k -> g || _ [[a|e|i|ĩ|o|u|ũ] [FLAGS]* [[a|e|i|ĩ|o|u|ũ]* [[c|k|t]|[FLAGS]] [a|e|i|ĩ|o|u|ũ]*]^>0 [FLAGS]* [c|k|t]]];*
2. define *PrenasalAugmentative {nd} -> t ||"+[Noun]" [ALLFLAGS]* ["+[Dim]" k a |"+[Augm]" k L | "+[Augm_Derog]" r S | "+[Augm_Coll]" m a | "+[Dim_Coll]" t S][ALLFLAGS]* "+[9NC]" [ALLFLAGS]* _ ;*

Other parts of speech implemented include adjectives, possessives, demonstratives, associatives, conjunctions, and prepositions. It is important to note that all but conjunctions and prepositions follow the concord system determined by the nouns they describe. Several analysis examples are shown below,

Nĩmakaamathaambagia
*1. +[Focus]+ [3P_PL]+ [Subj]+ [Rem_Future]+ [2NC]+ [Obj]**thaamba**+ [Verb]+ [DevExt_Causative]+ [Aspect_Habit]+ [FV_Ind]*
*2. +[Focus]+ [3P_PL]+ [Subj]+ [Rem_Future]+ [6NC]+ [Obj]**thaamba**+ [Verb]+ [DevExt_Causative]+ [Aspect_Habit]+ [FV_Ind]*
*3.+[Focus]+ [2NC]+ [Subj]+ [Rem_Future]+ [2NC]+ [Obj]**thaamba**+[Verb]+ [DevExt_Causative]+ [Aspect_Habit]+ [FV_Ind]*
*4.+[Focus]+ [2NC]+ [Subj]+ [Rem_Futsure]+ [6NC]+ [Obj]**thaamba**+ [Verb]+ [DevExt_Causative]+ [Aspect_Habit]+ [FV_Ind]*
*5. +[Focus]+ [2NC]+ [Subj]+ [Rem_Future]+ [3P_PL]+ [Obj]**thaamba**+ [Verb]+ [DevExt_Causative]+ [Aspect_Habit]+ [FV_Ind]*

kimagai
*1. **kima**+ [Verb]+ [Aspect_Habit]+ [FV_Ind]+ [Subjun_PL]*

gakaari
*1. +[Noun]+ [Dim]+ [9NC]**ngaari***

magoondu
*1. +[Noun]+ [Augm_Coll]+ [9NC]**ng'oondu***

tũguuka
*1. +[Noun]+[Dim_Coll]+[1NC]**guuka***

nyũmba
*1. +[Noun]+[9NC]**nyũmba***
*2. +[Noun]+[10NC]**nyũmba***

Word generation is carried much the same way as analysis, but in the reverse, as shown below;

+[Noun]+[10NC]mwana
1. ciana

4 Testing

In testing the morphology system we encounter a challenge as Gĩkũyũ, being a resource scarce language, has no existing standards against which this work can be evaluated against. We therefore adopt quantitative evaluation of the system's performance.

4.1 Morphology System

The system is evaluated on its efficiency in recognition of Gĩkũyũ words and correctness of analysis of test data. 100 words were randomly picked from the test data and analyzed using the morphological analysis system. The possible outputs from the evaluation point of view were;

i) Recognized and correctly analyzed,
ii) Recognized and incorrectly analyzed, and
iii) Not recognized.

5 Results

From the tests, it was observed that non-recognized words were mainly caused by the root form not being included in the lexicon files. Another category of non-recognized words was caused by the writers influence on spelling especially on vowel length and assimilation.

Result	**Correct Analysis**	**Incorrect Analysis**	**Not recognized**	
No. of instances	56	7	37	100
Precision	56/62 = 89.9%			
Recall	56/88 = 63.64%			
Success rate	56/100=56%			

Table 1: Morphology Results

6 Conclusion

In this work, we have explored the use of finite state methods to model morphological analysis of Gĩkũyũ, a resource-scarce Bantu language. Since Gĩkũyũ language is closely related to a number of Bantu languages, we propose the use of this knowledge and tools be applied to development of such languages.

Future work includes the application of the developed morphology system to implement a proof-of-principle shallow-transfer machine translation system for Gĩkũyũ to English.

Acknowledgments

The research work presented in this paper was made possible through the support provided by

Acacia Programme in IDRC, Canada through the PALnet/ANLoc Project.

Focus+Subj_Marker+Neg+Cond+Tense+Obj_Marker+Redupl+Verb+Dev_Ext+Aspect+FV

Figure 1: Verb Morphology

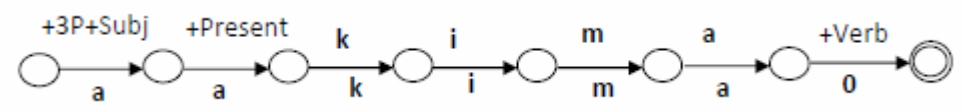

Figure 2: A Transducer Example

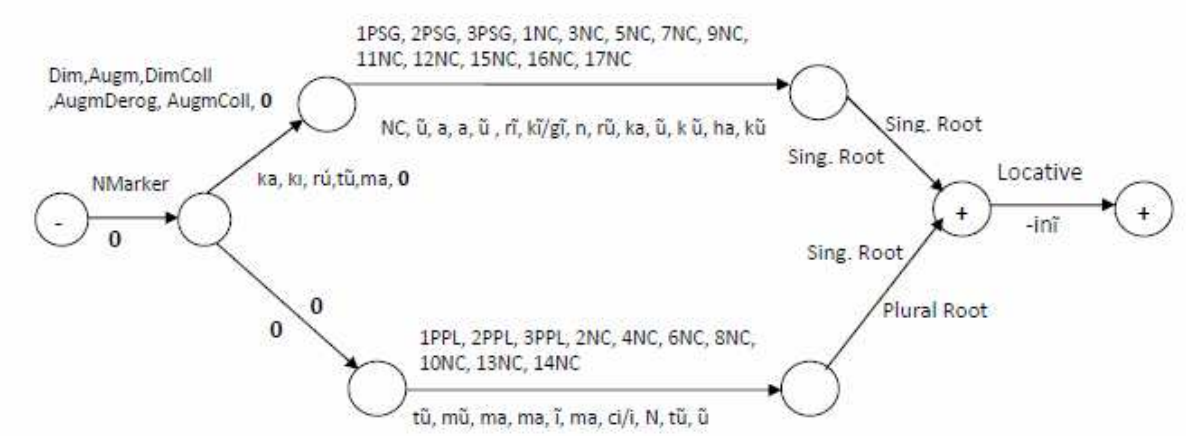

Figure 4: Underived Nouns

References

Beesley K. R. and Kartunen L., 2003, Finite-State Morphology, *CSLI Publications*, Stanford.

Dalrymple M., Liakata M., Mackie L., 2006, Tokenization and Morphological analysis of Malagasy, *Computational Linguistics and Chinese Language Processing*, Association for Computational Linguistics and Chinese Language Processing

G. De Pauw, Wagacha P., 2007, "Bootstrapping Morphological Analysis of Gĩkũyũ Using Unsupervised Maximum Entropy Learning". *Proceedings Eighth INTERSPEECH Conference.*

G. De Pauw, Wagacha P., De Schryver G. 2007, "Automatic Diacritic Restoration for Resource Scarce Languages". *Proceedings of Text, Speech and Dialogue, Tenth International Conference Heidelberg, Germany.*

http://www.aflat.org/?q=biblio *Publications on Natural Language Processing research* Papers on African Languages. (Accessed 12th March 2011)

Hurskainen A., 2004, HCS 2004 - Helsinki Corpus of Swahili. *Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC – Scientific Computing.*

Inaki A. et al, 2007, Shallow Transfer MT Engine for Romance Languages in Spain, *Universitat d'Alacant.*

Koskeniemmi K., 1983, Two-Level Morphology: A new Computational Model for Word-Form Recognition and Production, *University of Helsinki*.

Mugane J., 1997, A Paradigmatic Grammar of Gĩkũyũ *CLSI Publications*, Stanford California.