

# 1

---

## **Finite State Methods for Bantu Verb Morphology**

ROBERT ELWELL

### **1 Introduction**

This paper details the implementation of finite-state methods for the analysis of agglutinating verbs in the under-documented Bantu language Ekegusii. The transducer implements Xerox Finite-State Tool software (xfst) to arrive at a morpheme-by-morpheme gloss of a given verb, and conversely will return a grammatical verb given a morphosyntactically and semantically coherent sequence of glosses.

Dr. Lauri Karttunen, inventor of xfst, has declared finite-state morphological analysis a solved problem. (2006) He asserts that the morphology of any language can be analyzed using finite-state methods, and uses his xfst software to show the ease of implementation on Finnish, a language that several decades ago seemed impossible to account for computationally, in part due to long-distance dependencies that also impact the morphosyntax of Ekegusii. While his claim seems to have no clear counterpoint, utilization of finite-state methods, xfst in particular, will at the very least bolster this claim for every new language which is successfully implemented.

My goal was to utilize the speed and accuracy of finite-state methods in morphological analysis of the Bantu language Ekegusii, both to affirm claims made regarding the perspicacity of the usage of the finite-state transducer in morphological analysis, as well as to prove as an initial exercise for an expandable tool for Bantu language documentation and analysis. The script takes into account the major phonological processes of the language, accounting for tones and segmental sound alternations.

## 2 Language Overview

Ekegusii, an E. 10 Bantu language spoken in Western Kenya, has a rich agglutinating verbal morphology that can represent complex sentences monolexically. An overview of the morphological skeleton is as follows:

Finite:

(IN) - SM - TM - (LM) - (OM) - ROOT - (EXT) - FV

Non-Finite:

{(PPFX) / (IN) } - PFX - NEG - OM - ROOT - (EXT) - FV

Figure 2.1: The Morphological Skeletons for Ekegusii Verbs.

Abbreviations follow generally observed Bantu morphological positions. PPFx: Pre-Prefix, PFX: Class Prefix, IN: Initial Nasal, SM: Subject Marker, TM: Tense Marker, LM: Limitative Marker, OM: Object Marker, EXT: Extension, FV: Final Vowel.

Because of the large amount of morphemes found here, morphological analysis will be restricted to the verb, but will still represent a large amount of attested forms in the language. The following forms show the complexity of verbal glosses (Data from Elwell 2005):

n-to-ráá-minyok-e	na-to-mínyòk-á	okó-raager-a
IN-SM-TM-run-FV	IN-SM-run-FV	INF-'eat'-FV
'we may run'	'even if we run'	'to eat'
okó-raager-i-a	okó-raager-er-a	okó-raager-er-i-a
INF-'eat'-cause-FV	INF-'eat'-app-FV	INF-'eat'-app-cause-FV
'to feed' (cause to eat)	'to eat for (someone)'	'to feed for (someone)'
n-n-áa-ga-tam-íré		tw-áa-ka-minyoók-íré
IN-SM-TM-LM-'flee'-FV		SM-TM-LM-'run'-FV
'I never used to run away'		'we ran regularly'

Figure 2.2: Examples of Verbs

The verbs exhibited capture the capability of argument structure manipulation as well as the multi-faceted tense-aspect system within the verb.

## 3 Considerations for Application

Ekegusii has morphemic dependencies that must be considered to avoid creating unattested forms in the FST. These grammatical dependencies can stretch across several different morphemes, and are illustrated here:

<u>Element</u>	<u>Morphemes Involved</u>	<u>Agreement Influences:</u>
Aspect	(IN) TM, LM, FV	perfective, limitative, unmarked
Mood	SM, FV	negation, subjunctive, indicative, habitual
A-Structure	Root, OM, EXT	transitive, ditransitive, intransitive, etc.
Binding	SM, OM	reflexivity

Figure 3.1: List of Morpheme Dependencies

Beesley (1998) cites long-distance dependences as a “practical challenge” for finite-state morphology. Another point of attention is the phonological phenomena to be accounted for to arrive at the correct orthographic forms:

<u>Word</u>	<u>Gloss</u>	<u>Word</u>	<u>Gloss</u>
ogó-tam-a	‘to flee’	okó-raager-a	‘to eat’
px. - ‘flee’ - FV		px. - ‘eat’ - FV	

Figure 3.2: Dahl’s Law

An example of a prolific velar voicing dissimilation process in Bantu.

As in many Bantu languages, a closed class of verbal morphemes with a velar stop dissimilate with the voicing of the following onset (known as Dahl’s Law). This includes the infinitive marker as well as certain second-person morphemes. The alternation is captured with the following generative rule:

$$\begin{matrix} [-\text{son} & ] \\ [+hi & ] \end{matrix} \rightarrow [-\alpha \text{ voice}] / \text{ \_\_\_ } [+syl] + \begin{matrix} [-syl & ] \\ [\alpha \text{ voice} & ] \end{matrix}$$

Figure 3.3: A Generative Rule for Dahl’s Law in Ekegusii

Nasals in the syllabic peak or coda assimilate to the place of the following onset, adding to complexity in both analyzing and producing verbal forms:

n-raager-a	m-bún-á	ŋ-é-á
SM-‘eat’-FV	SM-‘break’-FV	SM-‘give’-FV
‘I eat’	‘I break’	‘I give’

Figure 3.4: Morpheme Alternation Due to Nasal Assimilation

Alternations of the first-person morpheme /ŋ/.

Once again, the generative rule for this process can be used to fashion a proper equivalent within xfst:

$$[+\text{nasal}] \rightarrow \begin{matrix} [\alpha \text{ cor} & ] \\ [\beta \text{ ant} & ] \\ [\gamma \text{ hi} & ] \\ [\delta \text{ bk} & ] \\ [\epsilon \text{ lo} & ] \end{matrix} / \text{ \_\_\_\_\_\_ } \begin{matrix} [-syl & ] \\ [\alpha \text{ cor} & ] \\ [\beta \text{ ant} & ] \\ [\gamma \text{ hi} & ] \\ [\delta \text{ bk} & ] \\ [\epsilon \text{ lo} & ] \end{matrix}$$

Figure 3.5: A Generative Rule for Homorganic Nasal Assimilation

Another phonological consideration, known as compensatory lengthening, restricts certain vowel clusters from coinciding, causing one vowel to become part of the onset, and the other to receive an extra mora.

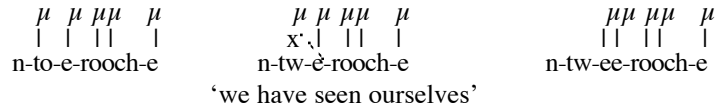


Figure 3.6: Compensatory Lengthening

In a process that is similarly motivated by the syllable peak, there is a restriction to more than two adjacent mora, and therefore syllable peaks themselves:

$$\begin{array}{l} \text{a-é-é-gér-a} \rightarrow \text{é-é-gér-a} \quad \text{‘he learns himself’} \\ \text{V} \rightarrow \emptyset / \_ \text{V V} \end{array}$$

Figure 3.7: Vowel Limitation and Its Subsequent Generative Rule

Finally, the tonology of the language must be accounted for. Like most Bantu languages, Ekegusii has two tones—high and low. High tones are treated as marked, and low tones are treated as default. As will be seen in the xfst execution, this observation is made clear by only marking high tones. There are three major rules to be accounted for in regards to tones: tonal spreading, melodic high docking of floating tones, and tonal shifting due to the obligatory contour principle.

Among toned segments, one of the most regular is the peak of the first syllable of the root, as roots can be divided into underlyingly toned and toneless (cf. Elwell, 2005), with different predictable phonological outputs. Among segments that are underlyingly high toned, it is apparent that a rightward tonal spreading occurs:

<u>U.R.</u>	<u>Surface Form</u>	<u>Gloss</u>
/m-bún-a/	[m-bún-á]	‘I break’
/n-tóm-a/	[n-tóm-á]	‘I send’

Figure 3.8: Tonal Spreading in Underlyingly Toned Roots

Specific tones license a floating high tone (FHT) which results in up to the last three syllables receiving a high tone. This is constrained by the Obligatory Contour Principle, which prevents floating high tones from attaching to peaks adjacent to toned syllables.

<u>FHT UR</u>	<u>Surface</u>	<u>UR (no FHT)</u>	<u>Surface</u>
/n-n-áá-bún-et-e/	n-n-áá-bún-eté	/n-n-áa-bún-et-e/	n-n-áa-bún-ét-é

Figure 3.9: ‘I broke (something)’: Two TAMs, one with FHT

As previously discussed, the Obligatory Contour Principle also has a great impact on the surface representation of high tones in Ekegusii, preventing spreading to syllables adjacent to toned syllables. A high tone will shift leftwards one syllable peak if it is next to a high tone. If it cannot shift leftwards, downstepping occurs. This is found phonetically as what can be considered a super-high peak preceding a high peak despite the fact that both tones involved are of the same height.

UR:	/o-kó-bún-a/	/kó-bún-a/
Shifting:	ó-ko-bún-á	_____
Downstepping:	_____	kó- <sup>1</sup> bún-á
Surface	[ó-ko-bún-á]	[kó- <sup>1</sup> bún-á]

Figure 3.10: Shifting versus Downstepping

A last consideration for tones is the habitual aspect morpheme, which is a high tone that anchors to the peak of the word-final syllable, resulting in an upstepping process. Usually, the tones of Ekegusii will result in a prosodic downdrift where the following high tone is actually lower than the high tone preceding it. However, in this case, following the analysis given in Elwell (2005), the word-final anchoring causes an unusual rise in tone relative to the downdrifting effect to be expected. The habitual aspect can be added to most tenses to varying degrees of semantic meaning.

n-áa-bún-été	n-áa-bún-etê
1sg-pst-‘break’-past	1sg-pst-‘break’-past-hab
‘We had broke’	‘We used to break regularly’

Figure 3.11: Habitual Tonal Morpheme (represented as circumflex)

In (3.11), the upstepped high tone is represented with a circumflex diacritic for convenience in application using UTF-8 encoding. For a more in-depth treatment of tonal phenomena in Ekegusii, see Bickmore (1997).

## 4 Previous Methods of Application in Bantu

Much work has been done in utilizing finite state methods for Bantu language processing. Using *xfst*, Karttunen (2003) uses a realizational framework to model Lingala verb morphology, taking a cue from the realizational morphology works of Anderson (1992) and Stump (2001). This approach focuses on using replacement rules to gradually construct the verb from the root, piece by piece. *xfst* has also been utilized for analyzing the Zulu verb (Pretorius & Bosch 2003). This analysis uses a *lexc* format for the morphological lexicon. Both approaches necessitated combinatory constraints to

allow for only semantically coherent forms. To capture long-distance dependencies, Karttunen builds these restrictions within the replacement rules, while Pretorius & Bosch utilize flag diacritics.

Work has also been done in Swahili using a language-specific morphological parser designed by Hurskainen (1992) known as SWATWOL. This parser is a two-level analyzer that similarly accounts for morphosyntax and morphophonology. SWATWOL accomplishes comparable results utilizing finite state methods, and is similar to an xfst script that utilizes *lexc*. It also makes use of diacritics to account for long-distance dependencies. SWATWOL has been used successfully in data and information retrieval using various text sources (Hurskainen 1995).

## 5 Implementation

A morphological skeleton approach was chosen in favor of a realizational approach. This was due to largely to my fieldwork (2005), which focused on description through this very framework, as opposed to theoretical syntactic analysis. Maintaining this is important for the marriage of the theoretical, descriptive, and computational sides of linguistics. Featural agreement also seems to be better suited by flag diacritics and unification, which captures the greater breadth of long-distance dependencies in Ekegusii.

Both finite and non-finite constructions were captured using a single regular expression. As for agreement, infinitive forms cannot have a subject; intransitive verbs cannot have an object; negative portmanteau subject morphemes must have negative final vowel endings, and the negative final vowel must agree in tense as well. The morphosyntax is asserted by defining every morpheme position as a set of its morphemes and their glosses. The morphemes are then configured into a regular expression as follows:

```
define ROOT[STEM (EXT) FV];
define WORD[(|[|(|(|(| IN)| PPFX) PFX) IIN)] SM TM (OM) (LM) ROOT];
```

Figure 5.1: Skeleton-Based Regular Expression

These accept a gloss and return a phonemic form of the verb, or vice versa. Unconstrained, the combinatory capabilities are vast, but result in many unattested forms. Because of the need to represent the many dependencies in Ekegusii, implementation in xfst necessitated several flag diacritics. Flag diacritics operate in a manner much like a unification grammar does in syntactic parsing—by allowing, disallowing, or requiring certain features to be present for the sake of grammatical agreement. In this case, diacritics are in conjunction with the presence of specific morphemes in combination within the FST. The following set of attribute-value pairs were utilized :

<u>Attribute</u>	<u>Values</u>	<u>Purpose</u>
TNS:	INF, PRES, RPAST, FPAST	Final vowel, TM, LM agreement
MOOD:	IND, SUBJ	Final vowel agreement
ASP:	PERF	Final vowel, TM agreement
NEG:	YES/NO	Final vowel, SM agreement
SUBJ:	1S, 2S, 3S, 1P, 2P, 3P	Reflexivity in OM
TRANS:	YES/NO	Argument structure, OM license

Figure 5.2: A List of Attribute-Value Pairs (cf. Fig 3.1)

The following example shows the use of flag diacritics in the script:

```
define TM [ “@U.TNS.FUT@” “@U.MOOD.SUBJ@” “D.ASP.PERF”
{+fut}:ráa | “@U.TNS.PAST@” “@U.ASP.PERF@” {+perfpast}:a];
define FV [ “@U.MOOD.SUBJ@” “@D.ASP.PERF@” {+subj}:e |
“@U.ASP.PERF@” {+perf}:ire ];
```

Figure 5.3: Examples of Flag Diacritics

Here, it is shown that in the FST known as TM, the tense marker for the future requires unification of that feature for the tense attribute and subjunctive feature for the mood attribute (a shortcut for final vowel agreement). Unification here, as in unification grammars, requires the correct attribute-value pair to be found in any case where a value is listed for the attribute after the diacritic is declared. As in unification grammars, violations of uniqueness result in ungrammatical forms. It can be seen here that attribute-value pairs are used to prevent both /ráa/ and /ire/ from co-occurring, as well as /a/ and /e/, while concurrently allowing the correct combinations of agreeing morphemes. First, there is the disallowance of specific attribute-value pairs (seen with the operator D preceding the pair). This prevents any item requiring unification of that pair or having that attribute value from co-occurring with an item that may have this rule upon it. Furthermore, unification, as previously stated, necessitates attribute-value exclusivity.

Using these constraints on the morphosyntactic interface creates structures that would be considered semantically coherent. However, verb extensions which can augment the argument structure (Mchombo 2005) may necessitate extraverbal information to create a completely coherent semantic form, preventing some cases of monolexical semantic completeness.

Phonemic forms can be pushed through a separate FST, returning a semi-orthographic representation. This FST is a composition of several phonological processes such as described above and more. Replacement rules are used as a close comparison to generative phonological rules:

```

define DISSIM [[k -> g || #.(o)_ V Cunvoi] .o. [g -> k || #.(o)_ V
Cvoi]];
define COMPLEN [[o -> we || ~V _ e ] .o. [o -> wa || ~V _ a] .o. [ ó ->
wá || ~V _ a] .o. [ ó -> wé || _e] .o. [o -> wá || ~V _ á] .o. [o -> wé || ~V
_ é]];
define NASSIM [[Cnas -> n || _ [tldlrslcln]] .o. [Cnas -> N || _ [kglN]]
.o. [Cnas -> m || _ [blm]]];

```

Figure 5.4: Replacement Rules (cf. Fig 3.4-6)

Phonemes not represented in the Latin character set are given capital letters in the FST and a final transducer composed into the phonology FST returns multi-character orthographic forms.

With the segmental phonology of the language accounted for, the tonal phonology must now be included. First, vowels that are underlyingly linked to high tones are represented in their morphemic regular expressions as their vowel counterparts with acute accents. This presents some amount of difficulty, because *xfst* (nor most computational applications) is not aware that, for example, [á] is the toned counterpart of [a], but rather that they are just two different characters. This, therefore, must be explicitly accounted for in any phonological rules to be articulated through replacement. However, by treating underlyingly high vowels in this manner, we retain the knowledge of what has a direct link to the tonal segment versus what is high-toned from spreading or other phonological processes. The only necessary step in ensuring this is treating vowels that become high-toned via spreading as marked in some intermediate form.

The approach taken in this FST is to take advantage of the language's insistence upon open syllables and use characters that do not represent a sound in the language as markers for tonal changes placed in the coda position. The steps taken to treat tonal spreading, shifting, and downstepping can all be seen via the following example:

UR:	/o-kó-bún-a/	/kó-bún-a/
Shifting 1:	o-kóz-bún-a	_____
Shifting 2:	oq-kóz-bún-a	_____
Downstepping:	_____	kó-!bún-a
Spreading:	oq-kóz-bún-aq	kó-!bún-aq
Removal:	oq-ko-bún-aq	_____
Attachment:	ó-ko-bún-á	kó-!bún-á
Surface:	ókobúná	kó!búná

Figure 5.5: Order of Replacement Rules to Account for Tonal Phonology  
The order given here reflects order of composition of replacement rules.

While using this marking-and-replacing strategy is perhaps less motivated than an autosegmental or optimality-theoretic treatment, the end result is essentially the same. Karttunen (2006) has asserted that despite the multiple possible treatments of phonology in xfst, the end result is invariably the same given the proper approach to the problem, ostensibly asserting xfst as a way to represent morphology, but not an actual model of the processes therein. This, then, is one way of modeling the patterns of the output of tones without focusing on modeling the process itself that underlies or motivates the patterns of the output. So while the actual movement taking place is motivated autosegmentally, it is represented in a loosely derivational format while keeping the end-product of the autosegmental analysis.

A final consideration involves two types of tones that seem to anchor to the right edge of the word. These are the floating high tone that exists in specific tenses and the tone that represents habitual aspect. While both of these tones must be treated as some non-orthographic markup character word-finally, time must be taken to ensure that the high tone that is linked to the right edge of the word actually appears in the correct manner without interference from the FHT. This is accounted for in two ways: first, the characters representing the floating high tone and the habitual high tone are different, preventing the upstepping rule from affecting a floating tone; second, the tonal regular expression FST that is found word finally is a disjunction of the floating high tones reflecting their corresponding tenses and the habitual high tone. At no time can these tones actually interact, and this is to be expected, because the upstepping actually can color the tonology of the word all the way to the antepenultimate syllable, which is as far into the word as the floating high tone can attach and spread rightward iteratively.

With both areas of the phonology accounted for and the lexicon constructed, the last step is composing these two modules in a manner that will have the gloss on one side and the phonologically-processed output on the other side. Composition of these two FSTs work in the following way:

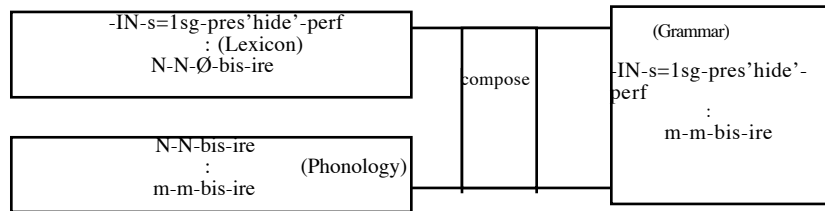


Figure 5.: A Graphical Representation of Composition

The Lexicon is actually an ordered regular expression comprised of several modular regular expressions while the phonology is a series of replacement rules composed into the proper order.

## 6 Results

The FST has on its upper end glosses that represent the basic features of every morpheme. On its lower end are attested semi-orthographic forms, as orthographic forms do not display tonal information. The following is an example of the Ekegusii morphological analysis in the xfst format:

```
xfst[2]: up ngók!ágeranire
-IN-inf+lim'weigh'-rec-perf
xfst[2]: down -IN-inf+lim'weigh'-rec-perf
ngók!ágeranire
```

Figure 6.1: xfst Execution

‘And then to have weighed for’ both analyzed and produced.

## 7 Further Considerations

With the FST complete and the proper forms accounted for phonologically and combinatorially, time could be taken at this point to improve upon the grammar by adding more roots and following what results could be found. By widening the range of verbal roots, one problem that will be encountered is a phonological process known as imbrication that affects some segments when they are in proximity to each other between the end of the root and the rest of the stem. This phenomenon is epidemic in Bantu languages, and while it is easy to recognize and motivate, it is difficult to account for. The best way to handle this problem is an optimality-theoretic approach, as derivationally the treatment is a convoluted series of rules with a large amount of exceptions. Regardless of either treatment, imbrication could be treated as a lexical phenomena, but this would require adding irregular entries to the ultimate FST agreeing in certain features but related somehow to the verb stem. Because of the limited amount of verbs that this affects, it is perhaps more worthwhile to recognize the strength of finite-state application for the vast majority of verbal forms, and possibly use statistical methods for learning imbricated forms.

## 8 Future Applications

An ultimate plan for the script will eventually include the ability to enter Ekegusii roots and their glosses through a python script and add them to the xfst lexicon. This final stage will transform the FST from an exercise in

morphological analysis to a tool for field work, linguistic analysis, and weak translation. Further considerations for future applications include a tool for Bantu cross-linguistic field work, composing glosses of the lingua franca Swahili with the glosses and morphosyntactic rules of other Bantu languages to arrive at a form, or for formal analysis through verbal comparison.

## 9 Conclusion

With Ekegusii verbal morphology properly accounted for, finite-state morphological analysis has again proven successful. The initial decision to stay focused on the regular expression represented a trade-off between specific theoretical frameworks and computational speed. However, by utilizing the regular expression for the combination of morphemes as well as lexical item categorization, a greater amount of readability of the script itself is accomplished. It seems that in the future, while *lexc* may be necessary to account for large amounts of new root forms, because the other morpheme categories are, in fact, closed classes, it may be possible to keep much in its current state.

This has been a first step in what has as of late been a growing trend: the utilization of computational methods for many uses related to under-represented indigenous languages, field work being one of these. This FST paves the way for computational applications in Ekegusii that could improve the information literacy and global connectedness of a speaker group of more than a million people, given the proliferation of computer-based information retrieval and the growing ubiquity of the Internet.

Theoretically, this exercise practically asserts a concept that I have also been examining using the LFG framework, focusing on the final vowel. Both formalisms, *xfst* and LFG, require featural constraints for the appearance of long-distance dependencies. The attested forms assert that there are abstract agreement features at play which influence combinatory possibilities. Furthermore, by theoretically motivating the constraining features in the lexicon, and using featural constraints in practice, a bridge has built between theory and application for verbal morphology and agreement in Bantu.

## References

- Anderson, S. 1992. *A-Morphous Morphology*. Cambridge: Cambridge University Press.
- Beesley, K.R. 1998. "Constraining Separated Morphotactic Dependencies in Finite State Grammars." *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing*: 41-49.

- Bickmore, L. 1997. "Problems in Constraining High Tone Spread in Ekegusii." *Lingua* 102:4, 265-290.
- Bresnan, J. 1987, With Sam Mchombo: "Topic, Pronoun, and Agreement in Chichewa." *Language* 63-4:741-782.
- Elwell, Robert. 2005. "A Morphosyntactic Analysis of the Ekegusii Verb." Honors Thesis Presentation, University at Albany, May 2005.
- Hurskainen, Arvi. 1992. "A Two-Level Computer Formalism for the Analysis of Bantu Morphology: An Application to Swahili." *Nordic Journal of African Studies* 1-1: 87-119.
- \_\_\_\_\_. 1995. "Informational Retrieval and Two-Directional Word Formation." *Nordic Journal of African Studies*. 4-2: 81-92.
- Karttunen, Lauri. 2006. "Finnish Numerals." Presented at University of Texas, February 24, 2006.
- \_\_\_\_\_. 2003. "Computing with Realizational Morphology." *Computational Linguistics and Intelligent Text Processing*. Alexander Gulbekh (ed.), Lecture Notes in Computer Science, 2588: 205-216.
- Mchombo, Sam. 2006. "Argument Binding and Morphology in Chichewa." Presented at Texas Linguistic Society 9, University of Texas at Austin, November 4-6.
- Pretorius, Laurette. 2003. With Sonja Bosch: "Finite-State Computational Morphology: An Analyzer Prototype For Zulu." *Machine Translation* 18:191-212.
- Stump, G. 2001. *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge: Cambridge University Press.