Chapter 11

Towards a unified theory of morphological productivity in the Bantu languages: A corpus analysis of nominalization patterns in Swahili

Nick Kloehn
University of Arizona

Models arguing for a connection between morphological productivity and relative morpheme frequency have focused on languages with relatively low average morpheme to word ratios. Typologically synthetic languages like Swahili which have relatively high average morpheme to word ratios present a challenge for such models. This study investigates the process of agentive nominalization from the perspective of the Dual Route Model. The findings suggest that all agentive nominal forms should decompose when accessed and thus that speakers of Swahili should include these morphemes in their lexical inventory apart from root morphemes. This process appears to not be influence by noun classification, or verbal derivation.

1 Introduction

Within the realm of MORPHOLOGICAL PROCESSING in LEXICAL ACCESS, a growing body of research has been conducted with the aim of developing a quantitative theory of MORPHOLOGICAL PRODUCTIVITY. The central notion of this body of research is the proposition that a morpheme's tendency to productively affix to a novel word-form is determined by whether existing words containing said affix undergo MORPHOLOGICAL DECOMPOSITION when accessed in the lexicon (Bybee 1995, Hay 2002, Hay & Baayen 2002). However, studies supporting these findings have been conducted in Indo-European languages which contain relatively low average morpheme per word ratios compared to the Bantu Languages.[1] This study aims to kick start a vein of research investigating the quantitative

[1] 2.55 morphemes per word for Swahili versus 1.68 morphemes per word for Modern English (Greenberg 1959).

Nick Kloehn. Towards a unified theory of morphological productivity in the Bantu languages: A corpus analysis of nominalization patterns in Swahili. In Jason Kandybowicz, Travis Major, Harold Torrence & Philip T. Duncan (eds.), *African linguistics on the prairie: Selected papers from the 45th Annual Conference on African Linguistics*, 155–168. Berlin: Language Science Press.
DOI:10.5281/zenodo.1251728

determinants of morpheme productivity in a group of languages which have higher average morpheme to word ratios than those previously studied. This first step is made by analyzing the process of nominalization in Swahili from the perspective of THE DUAL ROUTE MODEL (Baayen 1992) (Henceforth, DRM) in order to determine whether derived nominal morphology is predicted to be productive. The derivational process of nominalization is chosen by virtue of being analogous to the sorts of derivational patterns in English used as evidence to support DRM. However unlike English, nominalization in Swahili interacts both with noun classification and verbal derivation. This study touches upon the interaction of these morphological systems to illustrate the complexity of the issue of productivity in synthetic language. These systems are reviewed in §1.1 and §1.2. The study is contextualized with a discussion of the relevant Lexical Access literature in §1.3 as well as models of morphological productivity in §1.4 The history of morphological processing in synthetic languages, and the motivation for the current study is outline in §1.5. In §2 the paper outlines two corpus study designed to investigate the degree of predicted productivity of derived nominals in Swahili and the degree to which they interact with noun classification. In §3 the degree to whether or not productivity interacts with presence and number verbal extensions is evaluated. Finally, the results of these studies are discussed and future research is suggested in §4.

1.1 Nominal and verbal derivation in Swahili

One of the primary features common to all members of the Bantu language subgroup of the Niger-Congo language family is the presence of a rich noun class system (Heine 1982). In Swahili, noun class membership is indicated by a word initial prefix with little exception. These prefixes determine the grammatical number and semantic distinction of the nominal forms to which they are affixed. All nominals in Swahili must occur in some noun class even when the noun class is not signaled by a prefix. Regardless of whether the prefix is present, class membership can be determined by agreement patterns of agreeing prepositions, verbal inflection, nominal phrase constituents, and relativizers (Mohamed 2001). Example (1) gives the morphological breakdown for three nominals in Swahili each corresponding to a different noun class. The degree to which noun classification can be considered derivational or inflectional is debatable, but from the purposes of the present study it is only important to note that the process of noun classification has both inflectional properties (bound and obligatory) and derivational properties (changes meaning and category).

(1) (Swahili noun classes (Mohamed 2001, TUKI 2001))

a. mi-parachichi
CL4-avocados
'avocado trees'

b. vi-atu
CL11-shoe
'shoes'

c. u-nywele
cl8-hair
'hair'

Another morphological process that is present in Swahili is the system of verbal extensions. These suffixes denote a type of derivation that alters the semantic denotation of a verb and often the adicity of the verb undergoing this process.[2] Henceforth, a verb lacking a verbal extension will be described as simplex, and one which contains a verbal extension will be described as complex. Swahili takes advantage of its morphological system to represent semantically complex verbal environments in the form of morphologically complex verbs, whereas languages like English denote functionally analogous environments by both morphological and non-morphological means. For example, verbs that are semantically analogous in English to those derived through the process described in Example (2) must either (i) undergo the addition of derivational morphology (e.g. *lock - un-lock*), (ii) undergo the addition of inflectional morphology (e.g. *pay - paid*), (iii) introduce a phonologically and morphologically unrelated lexical item (e.g. *close - grasp*), or (iv) require a bi-clausal structure (e.g. *cook - make cook*). Furthermore, verbal extensions may occur multiple times forming multimorphemic complex verbs such as Causative-Applicatives.[3] Example (2) gives the derivation of two complex verbs (right of the arrow) and their simplex counterparts (left of arrow).

(2) (Complex verb derivations (Mohamed 2001, TUKI 2001, Seidl & Dimitriadis 2003))

a. yunj-a ↔ yunj-ik-a
break-FV break-STAT-FV
'break' 'be broken'

b. song-a ↔ song-o-a
press-FV press-AUG-FV
'press' 'press out'

Given that nominals in Swahili are mostly derived from verbs, the presence of verbal extensions affects the amount of morphology contained within many nominal forms. In other words, since nominals can be derived from both simplex and complex verb forms, the issue of whether a verb is simplex or complex is relevant to any investigation of nominalization. Regardless of whether a nominalized form is simplex or complex, the nominal will obligatorily be classified in a given noun class.

1.2 Nominalization in Swahili

In addition to the noun class prefix and optional verbal extension, the derivation of deverbal nouns requires the addition of one of three word final suffixes denoting nomi-

[2] I.e. some extensions (e.g. Causative) require the addition of an argument of the verb and others (e.g. Reciprocal) require the subtraction of an argument of the verb.

[3] E.g. *anz-il-isha* start-APPL-CAUS 'initiate'

nalization (Katamba 2003). In Swahili, nominalizing suffixes denote the relationship of the derived noun to the verb as either agentive or instrumental. Critically, agentive and instrumental forms are limited in the noun class prefixes which it may take,[4] as seen in Table 1.[5]

In addition, Table 1 shows that deverbal nouns may be formed from both simplex and complex verbs, and that they may also undergo both agentive and instrumental nominalization. Unfortunately, this paradigm does not cover the whole story. There also exist deverbal nouns for which there is a morphological variant suffix *-i*. [6] However, the current research limits the domain of inquiry to the suffix denoting agentive nominalization *-aji* which is loosely equivalent to the English suffix *-er* indicating the doer of an action.

1.3 Morphological processing in Lexical Access

The aim of the current subsection is to contextualize this study by reviewing the literature on how morphologically complex words are accessed. Research in the area of Lexical Access is able to rely upon the fact that frequent words are accessed in the lexicon faster than infrequent words. This WORD FREQUENCY EFFECT (Broadbent 1967) has been used to study morphological complexity as a methodology for determining whether affixed words are stored together or separately in the lexicon. Taft (1979) found in a VISUAL WORD RECOGNITION task that when multi-morphemic words of a constant frequency vary in the frequency of their stem morpheme, there was an asymmetry in time it took to access these words. Specifically, a word like size-d, which contains a stem of high frequency (*size*) was accessed faster than a word which has an identical surface frequency (*rake-d*), but which contain a less frequent stem (*rake*). This suggests, Taft concludes, that words containing a given stem are stored together in the lexicon (i.e. *size* is stored with *sized*, *sizable*, *sizing*, etc.). Similarly, it has been suggested that stems that occur with more non-inflectional affixes (e.g. *calculate: calculable*, *calculation*, *calculator*, *calculus*, *incalculable*, *incalculably*, *miscalculate*, *and miscalculation*) are accessed faster than those co-occurring with only few non-inflectional affixes (*roar: uproar*). This line of research has shown that reaction times in Visual Lexical Decision tasks are faster for nouns having a higher number of morphological neighbors for both mono-morphemic (Baayen et al. 2007, De Jong IV et al. 2000), and multi-morphemic nouns (Bertram et al. 2000)

Another vein of research has suggested that whether or not a complex word (e.g. *happiness*) is stored as one unit (*happiness*) or as multiple units (*happy+ness*) has to do with whether or not the component morphemes are parsed in perception (Baayen 1992 and Hay 2002). If one actively decomposes these forms during perception, then both units will be stored in the lexicon (as *happy* and *-ness*), and if one does not actively decompose them, then the form will be stored as a single unit (as *happiness*). Crucially, they claim that parsing is a function of the frequency ratio of the individual units. Specifically, parsing occurs when the frequency of the stem (*happy*) is greater than the frequency of the

[4] Class 3/4 may only occur with Instrumental Nominalization, and Class 1/2 may only occur with Agentive Nominalization.

[5] The associated verb is listed to the left.

[6] E.g. *mw-anz-il-ish-i* CL1-start-APPL-CAUS-AGENT_NOM 'founder'

Table 1: List of complex and simplex verbs and their corresponding deverbal nominal. These forms are divided by whether they are agentive or instrumentive nominals, and vary in their noun class. (Mohamed 2001,TUKI 2001, Katamba 2003).

Simplex Verb		**Deverbal Nouns**	
		Singular	Plural
cheza 'play'	Agentive Nominalization of Class 1/2	m-chez-aji CL1-*play*-AGENT_NOM 'player'	wa-chez-aji CL2-*play*-AGENT_NOM 'players'
	Instrumentive Nominalization of Class 3/4	m-chez-o CL3-*play*-INST_NOM 'game'	mi-chez-o CL4-*play*-INST_NOM 'games'
pepe 'fan'	Agentive Nominalization of Class 11	u-pepe-aji CL11-*fan*-AGENT_NOM 'fanning, waving'	—
	Instrumentive Nominalization of Class 11	u-pep-o CL11-*fan*-INST_NOM 'wind'	—
Complex Verb			
piga-na 'hit-RECIP'	Agentive Nominalization of Class 1/2	m-pig-an-aji CL1-*hit*-RECIP-AGENT_NOM 'fighter'	wa-pig-an-aji CL2-*hit*-RECIP-AGENT_NOM 'fighters'
	Instrumentive Nominalization of Class 3/4	m-pig-an-o CL3-*hit*-RECIP-INST_NOM 'battle'	mi-pig-an-o CL4-*hit*-RECIP-INST_NOM 'battles'
piga-na 'hit-RECIP'	Agentive Nominalization of Class 11	u-pig-an-aji CL11-*hit*-RECIP-AGENT_-NOM 'rivalry'	—
	Instrumentive Nominalization of Class 11	u-pig-an-o CL11-*hit*-RECIP-INST_NOM 'contest'	—

complex form (*happiness*), and does not occur when the ratio is the reverse. The model encompassing these concepts has been termed 'The Dual Route Model.' This tendency to parse has been argued to be hierarchically ordered from more separable units, which can freely attach to a base or stem (*-ness*), to less separable affixes which may only attach to a base (*-th*) (Hay & Plag 2004, Plag & Baayen 2009) pending semantic or syntactic selectional restrictions.

However, subsequent research has argued against this model by suggesting that morphological decomposition must occur in all cases (Taft 2004), that an information theoretical analysis will outperform the frequency accounts described above in predicting reaction times in a Visual Lexical Decision task (del Prado Martín et al. 2004) and that complex word-form frequency is a better predictor of decomposition overall (Baayen et al. 2007). So far, this research suggests that words are stored near their morphological neighbors, and are arguably stored in either morphologically complex or simplex forms.

1.4 Morphological productivity

The aim of this section is to describe the line that has been drawn from the status multimorphemic word storage (as one unit or as multiple units) to claims about morpheme productivity. Morpheme productivity is defined as the ability of a morpheme to occur in new environments, or to occur with novel words. For example, in English re- is qualitatively more productive than *en-*, because it can occur with words new to the language more readily (e.g. re-google: to google again versus #*engoogle*). Early research questioning the status of productivity has been qualitative (Schultink 1961) or as extremely difficult to measure (Aronoff 1976). However, newer models have attempted to categorize productivity as a consequence of word frequency. Bybee (1995) claims that token frequency (e.g. number of occurrences of *re+visit*) and type frequency (e.g. number of occurrences of *re*+STEM) are indicative of productivity. From this perspective, when a complex word-form has an individual token frequency that is relatively high, the word is more likely to be stored as a whole unit in one's mental lexicon and thus may not be decomposed. Words with a relatively low token frequency will tend to be novel, and if they are associated with a high type frequency then all of these words will be stored by each other in the lexicon. This storage, she argues, creates a stronger representation of the affix (e.g. *re-*). Bybee asserts that a stronger representation for a given morpheme should generally be associated with an increase in its productivity. Whereas Bybee's analysis does not make claims about a numeric threshold for this effect, Alegre & Gordon (1999) argue for a numeric threshold for a token frequency effect for complex forms above six per one million.

Baayen (1992) and Hay & Baayen (2002) claim that comparing token frequency (e.g. *re+visit*) and type frequency (e.g. *re*+STEM) alone is insufficient to account for productivity. Rather, the number of different decomposed forms acts as the predictor of productivity. They define decomposition as a function of the relationship between the derived form of a complex word (e.g. *re+visit*) and its underived or base (e.g. *visit*). If the derived form is lower than the base form, then the derived form will be parsed in perception. This

is measured by looking at the underived versus derived frequency of forms containing the affix in question (Hay 2002). If enough of these forms that contain the given affix (e.g. all words that contain the affix *re-*) are parsed in perception, then the morpheme will have an autonomous representation in the lexicon, and will therefore be productive. Here, there is a direct line between parsing and production: if an affix is parsed more often than memorized, it is stored separate from the words it affixes to, and is therefore predicted to be productive.

1.5 Rationale for present study

The research described has evaluated the relationship between frequency and productivity in languages like Dutch, English, French and Spanish (Baayen 1992, Bybee 1997, and Hay & Baayen 2002) but little work of this form has been conducted in more synthetic languages like Swahili. One such example is a study in Finnish which found that words which have larger numbers of morphologically related words elicit correct responses in a visual lexical decision task faster than those who had have smaller number of related words (del Prado Martín et al. 2004). The researchers claim that when a word in Finnish is morphologically related to many derived or complex words, those words tend to be accessed faster as opposed to the overall number of morphologically similar words in the language as is the case in Dutch. In another study on Finnish, Lehtonen et al. (2006) used fMRI to test neural activation during the recognition of complex words. They found that words which were morphologically complex tended to elicit activation in the region associated with semantic and syntactic processing, and that this might be evidence for morphological decomposition in perception.

Besides Finnish, other research has been done on languages with higher degrees of synthesis than previous languages including work on the productivity of nominal morphemes in Japanese. One such example is a study which tested both aphasic and non-aphasic speakers of Japanese on how they processed two nominal suffixes which varied in their productivity (Hagiwara et al. 1999). They claim that most adjectives in Japanese may derive a nominal by co-occurring with the highly productive *-sa* (e.g. *takai* → *takasa* versus *kebai* → *kebasa*), but only a subset of adjectives may derive a nominal by co-occurring with the less productive *-mi* (e.g. *takai* → *takami* versus *kebai* → **kebami*). They found that the processing of the two suffixes coincided in the activation of two areas: Broca's area for the productive (*-sa*) morpheme and the left-middle and inferior temporal areas for the semi-productive morpheme (*-mi*). To them, the difference in activation provides evidence for the claim that these two morphemes vary in a qualitative degree of productivity.

Overall, the work on typologically synthetic languages has suggested a difference in how these languages are processed compared to less synthetic languages. Whereas the results in the Japanese seem to support the predictions made by the DRM, nothing has been down to evaluate productivity in a language with the degree of synthesis in Swahili. The only experiment on Lexical Access in Swahili has argued that morphologically related words are stored near each other in the lexicon for second language speakers. Foote

et al. (2014) performed an experiment in Visual Masked Priming aimed at asking whether one can prime noun class prefixes within a semantic class (e.g. does *m-tu* 'CL1.SG-man' prime *wa-tu* 'CL2.PL-man'?) for L2 Swahili speakers. They found that when the primes were related by noun class (e.g. *vi-tanda* 'CL8.PL-bed'), there was an accelerated reaction time in accurately identifying the target word (e.g. *kitanda* 'CL.7-SG-bed'), but not to the same extent as to when the prime and target were identical (e.g. *kitanda - kitanda*). This suggests that words in the same noun class are stored in the same area in the mental lexicon for L2 speakers of Swahili. However, these findings are only a small step towards evaluating the relationship between productivity and word storage in Swahili. Therefore, this study evaluates the relative frequencies of nominal forms as outlined by the DRM in order to evaluate whether any asymmetry in productivity is predicted. This is performed by investigating the agent nominalizing affix *-aji* and the degree to which agentive nominalization interacts with the noun classification and verbal extension system. In order to see how often the suffix occurs (token frequency) and how many different word forms contain the suffix (type frequency) a frequency list is created from a corpus representing a sample population of Swahili words. In addition, token frequency is compared to type frequency and hypotheses are made to the likeliness of forms to be decomposed, and therefore the degree to which *-aji* may or may not be productive. Next, the log frequency of every instance of the morpheme (token derived frequency) is compared to the log underived frequency of each token in the same corpus (token underived frequency) in order to look at the overall trend in aggregate frequencies. Lastly, the number of morphemes per word is compared to the degree of predicted productivity of individual forms to determine whether a higher number of morphemes is correlated to a higher degree of predicted productivity.

2 Evaluating productivity

The aim of the current section is to describe a corpus analysis investigate that words that occur with the agentive nominalizing suffix *-aji*, and to answer whether the process can be described as productive based upon the models described above (Bybee 1995, Hay & Baayen 2002). The corpus study was performed using the the Helsinki Swahili Corpus of approximately 12.6 million words (HCS 2004) compiled from 12 news sources. The corpus of modern Swahili was used as a sample of the synchronic language present in speakers minds such that frequencies of word forms listed may very well reflect the frequency of occurrence of words contained in a speaker's lexicon. This commonly made behavioral inference leads us to the idea that frequency reflects the degree to which representations of a given word form has been accessed, and therefore will lead us to the idea ultimately of how productive these forms may be. In order to get around the issue as to whether complex verbs could be considered underived forms, all forms were included and controlled for in both analyses.[7]

[7] This issue could be a paper in itself. Hay & Baayen (2002) consciously limit their analysis to root forms and root+1 morpheme forms exactly because multiple affixation conflates the effect of relative frequency. Essentially, measuring the effect of underived versus derived frequencies get complicated when you have multiple affixation which themselves can denote both a derived and underived form relative to each other.

2.1 Measure 1: Token versus type frequency

The current subsection describes a Token versus Type frequency analysis of agentive deverbal nouns in Swahili. This was performed by using a regular expression to acquire all forms that ended in *-aji* along with the token frequency of each form. The regular expression used to obtain these words was unrestricted in order to include every possible token occurrence, and thus the resulting list required editing in order to filter out words that contained the word sequence *-aji* but were not related to the derivational suffix (e.g. *haji* 'pilgrim'; *jaji* 'judge') as well as words that were misspelled (e.g. **nitakusemaji* cf. *nitakusemaje* 'how I will tell you'). In order to obtain the type frequency (all occurrences of *-aji*), the token frequencies for each of the findings were summed. Each item was manually coded for whether it was simplex of complex,[8] and for its noun class. The resulting list was then analyzed in R, and a frequency table was made combining Noun Classes which vary only in number. Table 2 depicts the finding per Noun Class.

The data depicted in Table 2 demonstrate multiple important aspects of agentive nominalization in Swahili. First, there is a high tendency for nominalized forms to occur in 2 semantic fields (i.e. class 1/2 and class 11). Since class 1/2 denotes animates, and class 11 denotes abstract concepts the data show that there is a strong association of *-aji* with animacy, and with the abstract class. Second, the Token Frequency for each Noun Class (save for Class 5/6) is above the 6 per million threshold argued by Alegre & Gordon (1999) to be the minimum frequency for the parsing of complex forms. This indicated that for most of these forms, we would predict that the frequencies are not too low to exhibit

[8]Complex is divided into two groups: 1 verbal extension present or 2 verbal extensions present.

Table 2: Agentive deverbal nouns. Number of occurrences per million of tokens of the type STEM-aji in the Helsinki Swahili Corpus of 12,610,158 tokens. These tokens are divided by noun class ranging from the highest frequency to the lowest frequency occurrences per class. The Type Frequency column depicts how many unique words occur in each class per million, and the Token Frequency column depicts the raw number of words occurring in each class per million (given that many words have repeat occurrences this number is higher than the type frequency). The percentages highlight the proportion of total tokens that occur in a given noun class. The third columns shows that ratio of Type to Token occurrences per noun class.

Noun class	Type frequency (per million)		Token frequency (per million)		Type:Token
m-/wa- (1/2)	24.11	*47.28%*	876.03	*57.32%*	*0.0275*
u- (11/14)	23.39	*45.87%*	611.80	*40.04%*	*0.0382*
ki-/vi- (7/8)	1.98	*3.88%*	29.90	*1.96%*	*0.0662*
ji-/ma- (5/6)	0.95	*1.86%*	4.20	*0.27%*	*0.2262*
n-/n- (9/10)	0.56	*1.11%*	6.26	*0.41%*	*0.0895*
Total	50.99		1528.19		

parsing in perception. Lastly, and most relevant to productivity: the ratio of token and type frequencies are very similar between noun classes. A fair assertion to make here is that, given that the ratios are not wildly different, we would predict of similar degree of productivity across the classes if productivity is in fact a function of Type versus Token frequency. This would indicate that the tendency for tokens to occur in only two classes is a likely a semantic effect and not an effect of productivity.[9]

In order to determine whether productivity is predicted to occur equally in each noun class, a one-way ANOVA was used to test whether there has a correlation between noun class and the Type: Token ratio. Significance would suggest that if productivity were truly a function of Type:Token ratio, then we would predict asymmetries in productivity between noun classes, and non-significance would predict no productivity asymmetries between noun classes. Type:Token ratio did not significantly vary across the noun classes $F(1,574) = .767$, $p < .35$. This may indicate that agentive deverbal nominals will be equally productive across the noun classes, and that asymmetry in the overall frequency of these classes is conditioned by semantic restriction.

Collapsing across the classes, we can assess productivity from the perspective of Token versus Type frequencies by identifying the average number of occurrences for each type. According to Bybee, a productive morpheme would have many tokens that occur only a few times, as opposed to a few forms that occur many times. Figure 1 represents the token frequency for each Type as a histogram.

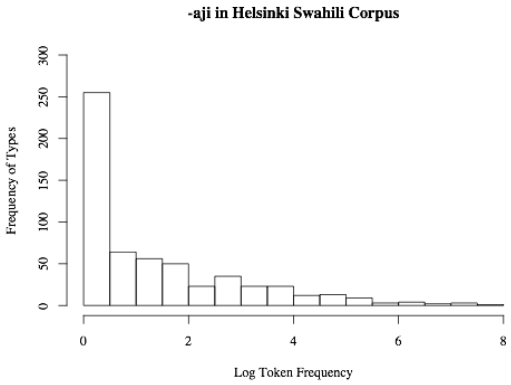

Figure 1: Histogram depicting the log of the Token Frequency along the x-axis, and the Frequency of Types occurring per log token frequency on the y-axis. This means that the bar furthest left depicts a group in which there are 250 tokens with a log frequency of zero (i.e. a frequency of 1, given that Log(1)=0).

[9]This notion stems from the idea that these noun classes are both semantically and functionally determined. Whereas class 1/2 and 11 have strong semantic distinctions, many others do not indicate semantic fields that are as consistently clear. In addition, most new words are put into class 9/10 (often unmarked) indicating a non-semantic classification, based on grammatical necessity of noun class membership in the language.

The histogram depicts data much like what Bybee's productive morpheme would resemble. There are around 250 type types with a log frequency occurrence of 0, moving right, there are over 400 types with a Log Token Frequency of less than 2, around 520 less than 4, and possibly only 50 types greater than 4. The result is clear: the majority of types have a relatively low token frequency, and only a few outliers represent the bulk of token frequency (more than 10 tokens).[10] The data show that one would predict productivity from the standpoint of a model that draws upon productivity from Token versus Type frequency analysis. If the relationship between Type and Token frequency is a cue for productivity, and not overall raw frequency, then one would predict a speaker to be equally accepting of a novel form in one noun class as opposed to another barring semantic ill-formedness.

2.2 Measure 2: Underived versus derived frequency

The aim of this subsection is to build on the analysis in the previous section in making a case for productivity by testing agentive deverbal nouns for the derived versus underived frequency. This was performed by using the list of all tokens from the previous analysis. A Perl regular expression was then used to strip all deverbal nouns of their noun class prefixes and agentive suffixes in order to find the underived forms. Given that verbs in Swahili may not occur without left peripheral inflectional prefixation, the underived words included verbs containing inflectional affixation.[11] No complete form of the Helsinki Swahili Corpus is openly available, and so frequency information may only be obtained by web query. Given the number of underived forms tested (574), the process was automated using a BASH cURL script which saved the returned token and frequency information in JAVA script. The results were concatenated into a single text file, and search and replace commands were used to transform the data into analyzable columns, to then be added to the data frame from the previous analysis.

In the vein Hay & Baayen (2002), the log frequencies of the underived (base) tokens were graphed against the log frequency of the derived tokens. Figure 2 demonstrates the relationship between base frequency (underived form) and derived frequency of each occurrence of *-aji* within the corpus. Within the DRM, whether or not a form is parsed is related to its position on this graph to an X=Y line. This is simply a line with a slope of 1, and a y-intercept of 0 that bisects points where the x and y measure are equivalent. In addition, they require that an r2 line describing the data should be significantly above the X=Y line in order for the morpheme to be productive. According to Hay & Baayen, this

[10] It is important to note here that forms that are derivationally related but vary in semantic noun class and not just number (e.g. word forms that occur in class 1 and 11) have not been combined into the same type category. An analysis combining the two or more types would gather together words that differ in in one derivational morpheme, but who share the same derivational root. Whether or not the frequency of one should effect the other will be left to future studies.

[11] Swahili verbs inflect as follows, with [] indicating obligatory inflection, and {} indicating optional inflection/derivation based upon the denotation of the proposition: [Subject]-[Tense/Aspect]-{Object}-Verb Stem-{Verbal Extension(s)}-Final Vowel *Chakula ki-na-ku-pik-i-w-a.* CL7.food CL7.AGR-PRES-2SG.OBJ-cook-APPL-PASS-FV 'The food is being cooked for you.'

relationship reveals whether we can predict that the morpheme in question has some sort of mental representation, and therefore occurs productively as seen in Figure 2.

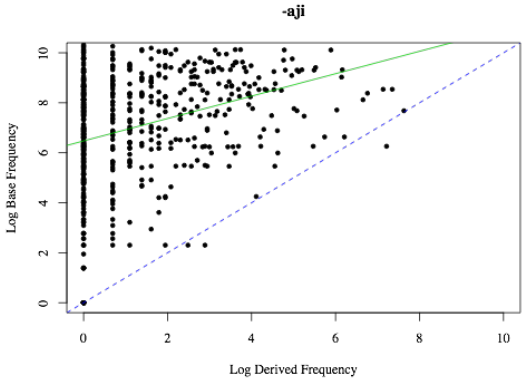

Figure 2: Log Derived Frequency pitted against Log Base Frequency. The blue dotted line represents that X=Y line. A data point on this line will have an equivalent derived and underived frequency. Points above this line have a higher Log Base Frequency than a Log Derived Frequency, and ones below have a higher Derived Frequency than Base Frequency. The green solid line represents the r2 line ($p > .0001$) which demonstrates the significant positive correlation between Base Frequency and Derived Frequency.

The Figure above shows data points that relate a single forms derived and underived frequencies. A data point below the line will not be parsed in perception according to the theory, and one above the line would be parsed. Clearly the bulk of the data is above the X=Y which means we would predict parsing in perception. The r^2 line describes this phenomenon mathematically. The relationship between the X=Y line and the r^2 line is significantly above the X=Y line, we can posit that the bulk of data points have a higher underived frequency than derived frequency. It is precisely this relationship that denotes productivity in the Dual Route Model.

The aim of this section was to strengthen a claim for productivity of the agentive nominal from the perspective of a dual processing route model. Both theories predict productivity of the affix. The next section aims to expand on this claim by analyzing morphological complexity within the nominal forms.

3 Morphological complexity and productivity

The aim of the current section is to analyze the relationship of the verbal extension system to the degree of how productive a given form is. We would predict that the presence of verbal extensions impacts productivity, then there will be an asymmetry in the predicted productivity of verbs containing verbal extension and those which do not. To do

this, the residuals of the r^2 line from §2 (the distance from the line describing the distribution) are compared to the number of verbal extensions present in the forms.

3.1 Regression analysis

In order to test this relationship, morphological complexity was tested against the degree morphological productivity measure using the residuals. A residual of zero means that the data point is on the r^2 line, and points above and below are indicated by positive and negative integers respectively. The assertion here is that points with a negative residual, while still being parsed, are working against productivity, whereas a positive residual is working towards a higher level of productivity. The question then is, is there any correlation between this tendency and whether or not there is a verbal extension present?

3.2 Results

A one-way ANOVA was used to test whether morphological complexity influenced distance for the r^2 line. Given the unevenness of scores per group (300 in Simplex, 150 in 1 Complex, and 59 in 2 Complex) a random sample of data was taken from the two larger groups of equal number to the smallest group. The groups Distance from the r2 line do not vary significantly across the three groups, $F(1, 58) = 26.89$, $p > .2$. However, as the box plot in Figure 3 shows, the non-truncated data shows a trend toward significance in which there is an inverse relationship between morphological complexity and distance from the r2 line. However, a nonsignificant affect suggests that the predicted productivity of denominal verbs is not impacted by whether or not verbal extensions are present.

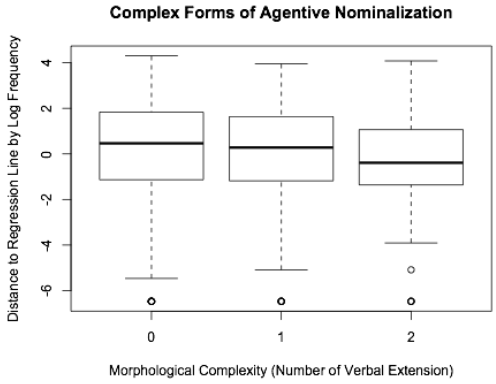

Figure 3: Box plot showing the three groups of morphological complexity on the x-axis (0 Verbal Extensions, 1 Verbal Extension, and 2 Verbal Extensions), and the distance to the regression line on the y-axis. The medians of the groups nonsignificantly decrease as the number of verbal extensions increase.

4 General discussion

Under each analysis given, one would predict that the affix *-aji* should be productive for both complex and simplex verb forms, and regardless of noun classification. The next question to be answered is whether or not all nominalized forms exhibit the same variation. For example, does the instrumental form *-o* exhibit the same degree of predicted productivity with no inhibition for the other morphological systems? A related question would be to isolate whether there are any morphological processes in Swahili that are non-productive. A corpus study on a much larger scale will be able to evaluate this by compare derived and underived forms of all verbs and nominals. If it is the case that these affixes are all predicted to be productive, then there is a testable prediction for a behavioral study investigating the ways in which Swahili speakers process complex words. Namely, the DRM would predict that all forms should be decomposed and therefore that there should be no significant asymmetry in reaction times between complex word forms beyond their base frequencies. The opposite result would indicate that speakers of Swahili may have different thresholds for parsing than speakers of more isolating languages like English. Such a finding would provide evidence for the idea that Lexical Access is ordered relative to the language you acquire. This would entail that some languages would make use of one route of processing more commonly than a language like English, or that these models are wide of the mark, demanding a reanalysis of the way in which human language is processed cross-linguistically.

Acknowledgements

The data in this paper were presented at The University of Arizona Graduate Student Showcase Feb 2014, and ACAL 45 at the University of Kansas. Thanks to all involved and to Mike Hammond, Heidi Harley, Adam Ussishkin and Andy Wedel for comments.

Abbreviations

Numbers in glosses indicate noun class prefixes and pre-prefixes. Abbreviations follow Leipzig glossing conventions, with the following exceptions:

AGENT_NOM	agentive nominalizer	FV	final vowel
AUG	augment	STAT	stative

References

Alegre, Maria & Peter Gordon. 1999. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40(1). 41–61.

Aronoff, Mark. 1976. Word formation in generative grammar. *Linguistic Inquiry Monographs* 1. 1–134.

Baayen, Harald R. 1992. Quantitative aspects of morphological productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1991*, 109–149. Netherlands: Springer.

Baayen, R. Harald, Lee H. Wurm & Joanna Aycock. 2007. Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon* 2(3). 419–463.

Bertram, Raymond, R. Harald Baayen & Robert Schreuder. 2000. Effects of family size for complex words. *Journal of memory and language* 42(3). 390–405.

Broadbent, Donald E. 1967. Word-frequency effect and response bias. *Psychological review* 74(1). 1.

Bybee, Joan L. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5). 425–455.

Bybee, Joan L. 1997. Semantic aspects of morphological typology. In Joan L. Bybee, John Haiman & Sandra A. Thompson (eds.), *Essays on language function and language type: Dedicated to t. Givón*, 25–37. Amsterdam: John Benjamins Publishing.

De Jong IV, Nivja H., Robert Schreuder & Harald R. Baayen. 2000. The morphological family size effect and morphology. *Language and Cognitive Processes* 15(4-5). 329–365.

del Prado Martín, Fermín Moscoso, Aleksandar Kostić & Harald R. Baayen. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition* 94(1). 1–18.

Foote, Rebecca, Patti Spinner & Rose Mwasekaga. 2014. *Morphological decomposition in Swahili. Presentation at Linguistic Society of America 2014 Annual Meeting.*

Greenberg, Joseph H. 1959. Africa as a linguistic area. In William R. Bascom & Melville J. Herskovits (eds.), *Continuity and change in African cultures*, 15–27. Chicago: University of Chicago Press.

Hagiwara, Hiroko, Yoko Sugioka, Takane Ito, Mitsuru Kawamura & Jun I. Shiota. 1999. Neurolinguistic evidence for rule-based nominal suffixation. *Language* 75(4). 739–763.

Hay, Jennifer. 2002. From speech perception to morphology: Affix ordering revisited. *Language* 78(3). 527–555.

Hay, Jennifer & Harald R. Baayen. 2002. Parsing and productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 2001*, 203–235. Netherlands: Springer.

Hay, Jennifer & Ingo Plag. 2004. What constrains possible suffix combinations? On the interaction of grammatical and processing restrictions in derivational morphology. *Natural Language and Linguistic Theory* 22(3). 565–596.

Heine, Bernd. 1982. African noun class systems. *Apprehension: das sprachliche Erfassen von Gegenstanden* 1. 189–216.

Katamba, Francis X. 2003. Bantu nominal morphology. In Derek Nurse & Gerard Phillippson (eds.), *The Bantu languages*, 103–120. London: Routledge.

Lehtonen, Minna, Victor A. Vorobyev, Kenneth Hugdahl, Terhi Tuokkola & Matti Laine. 2006. Neural correlates of morphological decomposition in a morphologically rich language: An fMRI study. *Brain and Language* 98(2). 182–193.

Mohamed, Mohamed Abdulla. 2001. *Modern Swahili grammar*. Nairobi: East African Publishers.

Plag, Ingo & Harald R. Baayen. 2009. Suffix ordering and morphological processing. *Language* 85(1). 109–152.
Schultink, Henk. 1961. Produktiviteit als morfologisch fenomeen. *Forum der letteren* 2. 110–125.
Seidl, Amanda & Alexis Dimitriadis. 2003. Statives and reciprocal morphology in Swahili. *Typologie des langues d'Afrique et universaux de la grammaire* 1. 239–284.
Taft, Marcus. 1979. Recognition of affixed words and the word frequency effect. *Memory & Cognition* 7(4). 263–272.
Taft, Marcus. 2004. Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology Section A* 57(4). 745–765.
TUKI, The Institute of Swahili Research at the University of Dar es Salaam. 2001. *English - swahili dictionary - kamusi ya kiingerza-kiswahili.* Dar es Salaam: Laurier Books Ltd. ISBN 9976911297.