1 Morphological Productivity in Swahili

Aim:

The aim of this paper is to evaluate whether affix productivity can be predicted by frequency ratios between derived and underived morphological forms in Swahili.

Hypothesis:

Previously the relationship between derived and underived forms have been argued to predict the productivity of English derivational affixes. We evaluate this relationship in Swahili. However, given Swahili's rich morphological system, we must redefine the notion of an 'underived' form. We redefine this notion by classifying underived forms as the root frequency without the affix. We expect that the ratio of our newly defined underived and derived forms should correlate with the number of hapaxes, or novel occurrences in the language.

Data:

We look at both nominal and verbal derivational affixes (all suffixes), as well as prefixed verbal inflectional and nominal forms. Data comes from the 13.2 million token Helsinki Corpus of Swahili.

Basic methodology:

For each affix in question, we look at all word forms containing the token, and count them. We then count the frequency of the root. All underived and derived forms are plotted in log space, and their productivity values are extracted from the plot, including the ratio of types that have greater frequencies in their underived forms than derived forms, and the slope and y-intercept of a best fit line. This is done for all affixes in 7 subsections of the corpus. Based upon these variables, we evaluate the degree to which they correlate with the number of hapaxes in non-overlapping sections of the corpus.

Results:

We find that our frequency ratios are overall positively correlated with the number of hapaxes in a non-overlapping subsection for all variables. This suggests that speakers of Swahili may rely upon root frequency in the way the English speakers rely upon underived forms [1].

[1] There is also a component to this paper looking at code-switching on Twitter, and whether these ratios relate to that. I as of yet have not found an effect

2 Logic

This is a description of the program pipeline for the paper on Swahili morphology to be submitted to Morphology. There are two Corpus project and 1 analysis project comparing the two:

- The CS Corpus containes a masterscript (propogate corpus.py) to call the codeswitching classifier (CS classifier.py) trained using a language model (is language.py) generated from some files in Corpus Resources (eng.txt, other.txt). The classifier itself trained on the other files in that folder (test.txt, and tweets.txt). In addition, the file Twitter API contains that script used to capture the tweets.

- HSC Corpus contains HSC.py which is uploaded to the server of the HSC corpus and extracts the morphological types, then generates underived forms from those types. It gets the counts of those types for each morpheme listed. The output is a string of .txt files of derived and underived forms for each morpheme. Those countes (of both derived and underived forms) are in count.txt. The Development files simply the working forms of the files captured from the corpus.

3 Analysis

analyze CS counts the token and types for each morhpeme in the CS corpus. HSC predictors analyzes the underived and derived froms from the HSC corpus

It has been argued that the gradient in productivity of a given morphological type correlates to how often words of that type are parsed in perception (Hay 2003, Hay and Baayen 2002). Whereas such a claim seems to hold in a language with an impoverished concatenative morphology like English, languages with greater morphological complexity present a challenge. This issue stems from the fact that morphological types in English have underived or bare forms, but for many languages, there is no such thing as a bare forms. In this paper we analyze Swahili verbal and nominal derivation, for which there are no such underived forms, and for which the models above would predict that no inflectional form should be parsible and therefore productive, in their current form. However, we know this not to be the case, and on the contrary, suggest that inflectional forms should nearly always be productive except for in cases of defectiveness (Sims 2014).

Additionally, Swahili derivational affixation is impossible without inflectional marking (1), so for a given derivational type there may exist multiple inflectional forms possible. This means, that when we look at derived and underived frequencies of inflected forms, there are two problems. First, if we look at a given morphological type, there is simply no such thing as a stem in isolation. Second, when we do look at derived forms, we are looking at a derivational type subdivided into multiple inflectional classes. While much has been said about the relationship between stackable derivational affixes (Siegel 1974, Kiparsky

1982, Hay and Plag 2009, Sims and Parker 2015), the issue here is that forms of a single derivational type must co-occur with another inflectional type, and these types may vary.

To get around this we work with the hypothesis that for these affixes, the frequency of morphological competitors (known as the cumulative root frequency (Cole et al. 1989)) is key in predicting productivity. This hypothesis implicitly alters our morphological model to suggest that productivity is not a result of discrimination between an affixed form and its affixless counterpart, but descrimination between an affix and all of its competitors of which the affixless form is just one possibility. In our corpus investigation, we perform a k-folds analysis 7 random sections of the Helsinki Corpus of Swahili (13.2 million words) in which we compare the productivty measure from Hay and Baayen (2002) for 60 affixes to the number of hapaxes of the affix (V1, Baayen 1993) in a non-overlapping subsection of the corpus. The results demonstrate that these two variables are significantly correlated, suggesting that descrimination of competing forms, and not un-derived forms is the key to predicting productivity in languages lacking non-derived forms.