Michel Généreux & Tjerk Hagemeijer (Lisbon University)

Experiments with the basic vocabulary of the Golf of Guinea Creoles

This paper sets out by measuring the orthographic and phonetic distance between the basic vocabulary (extended Swadesh list) in the four Portuguese-related Gulf of Guinea creoles: Angolar (AN), Santome (ST), Principense (PR) and Fa d'Amb^o (FA). We expect that the differences between orthography and phonetics not to be significant, because the writing system is already based on phonetics and phonology. Using computational methods, we will attempt to reconstruct the basic vocabulary of the proto-language from which the four creoles descend.
Keywords: Gulf of Guinea creoles, lexical distance, lexical reconstruction

1. Comparing their orthography
We use dynamic programming to compute the edit-distance between two words. This is the minimum number of operations (deletion, substitution, insertion) required to align two words together and this is called the cost or the distance between the words. Note that we use a uniformized writing system for the four creoles. Average distance per word-pair We measured the distance between a Swadesh list of 216 terms and a reduced list of 163 terms without African items, which are especially common in Angolar. We found that the distance between languages is less when we used the reduced list and that AN, FA and PR creoles are closer to ST than to each other.

2. Comparing their phonetics
This time we used edit-distance and the same vocabularies (Swadesh and reduced) to compute phonetic distance but with a phonetic model and a larger set of operations (e.g. fortition, lenition, aphaeresis, prosthesis, apocope, etc.). Note that the differences between orthography and phonetics are not significant, because the writing system is already based on phonetics-phonology.
Average cost per word-pair Results are similar to orthographic distance, albeit distances are slightly more marked. We also found that AN-FA appears most distant orthographically and phonetically, while PR-ST appears least distant orthographically and phonetically.

3. Reconstructing proto-lexical items from the Swadesh list
We adapted a computer program (Oakes, 2000) which implements the "Comparative Method" described in Crowley and Bowern (1998).
Results We automatically reconstructed 79% of the Swadesh lexicon for the four creoles.

4. Future Work
_ Examine the list of frequent sound changes to see if new regular patterns apply to Gulf of Guinea creoles.
_ Use the distance measure to extract potential cognates from Gulf of Guinea creole corpora.
_ Evaluate the list of proto-words we reconstructed.

Crowley, Terry and Claire Bowern. 1998. An introduction to Historical Linguistics. Oxford University Press.

Oakes, Michael P.. 2000. Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. Journal of Quantitative Linguistics, 7(3): 233–243.