Revisiting the Swadesh Wordlist: How Long Should It Be

Authors : Feda Negesse

Abstract : One of the most important indicators of research quality is a good data - collection instrument that can yield reliable and valid data. The Swadesh wordlist has been used for more than half a century for collecting data in comparative and historical linguistics though arbitrariness is observed in its application and size. This research compare s the classification results of the 100 Swadesh wordlist with those of its subsets to determine if reducing the size of the wordlist impact s its effectiveness. In the comparison, the 100, 50 and 40 wordlists were used to compute lexical distances of 29 Cushitic and Semitic languages spoken in Ethiopia and neighbouring countries. Gabmap, a based application, was employed to compute the lexical distances and to divide the languages into related clusters. The study shows that the subsets are not as effective as the 100 wordlist in clustering languages into smaller subgroups but they are equally effective in di viding languages into bigger groups such as subfamilies. It is noted that the subsets may lead to an erroneous classification whereby unrelated languages by chance form a cluster which is not attested by a comparative study. The chance to get a wrong result is higher when the subsets are used to classify languages which are not closely related. Though a further study is still needed to settle the issues around the size of the Swadesh wordlist, this study indicates that the 50 and 40 wordlists cannot be recommended as reliable substitute s for the 100 wordlist under all circumstances. The choice seems to be determined by the objective of a researcher and the degree of affiliation among the languages to be classified.

Keywords : classification, Cushitic, Swadesh, wordlist
Conference Title : ICLLL 2016 : International Conference on Languages, Literature and Linguistics
Conference Location : Vienna, Austria
Conference Dates : June 16-17, 2016