BULL. AUSTRAL. MATH. SOC.
VOL. 11 (1974), 477-478.

62M10, 05C05

Numerical taxonomy for languages

Annette J. Dobson

Linguists and anthropologists try to make inferences about the relationships and history of languages within a family from the vocabularies. They count the number of cognate words in each pair of present-day languages and use these data to try to reconstruct the ancestry of the family.

Usually the reconstruction is based on a simplified model of language development in which each language splits off from an ancestor language and then evolves independently without borrowing words from other languages. There are two related problems. The first is to deduce the sequence of splits; this is analogous to the biological problem of classifying species according to their degree of similarity. The second problem is to estimate the dates at which the supposed splits occurred. Actually, it is only possible from these data to find the times of the splits relative to one another; to obtain chronological dates additional information would be needed.

This thesis is mainly concerned with estimating the separation times of the languages although some combinatorial results about the use of unrooted tree-shapes in numerical taxonomy are also included. To estimate the times it is necessary to postulate a stochastic model for the way that words which correspond to a particular meaning may, in time, be superseded by other words. It is usual to use some form of exponential distribution as a model for word replacement. Then a statistical procedure is needed to estimate the times and to characterise the distribution of word replacement.

During the last twenty years, various models together with appropriate

Received 7 August 1974. Thesis submitted to the James Cook University of North Queensland, March 1974. Degree approved, September 1974. Supervisor: Professor B.C. Rennie.

estimation procedures have been proposed; one of the most notable is by Dyen, James, and Cole [1]. The model discussed in the thesis and the estimation procedure are different from preceding ones. Properties of the estimators are investigated theoretically and by simulation. Then separation times are estimated from two sets of real data, one for some Indo-European languages and the other from the South Pacific area.

Reference

[1] Isidore Dyen, A.T. James, and J.W.L. Cole, "Language divergence and estimated word retention rate", *Language* 43 (1967), 150-171.