Proc. Indian Acad. Sci., Vol. 86 A, No. 6, December 1977, pp. 503-512, © Printed in India.

Automatic phonetic transcription of Tamil in Roman script

E V KRISHNAMURTHY
Department of Applied Mathematics, Indian Institute of Science, Bangalore 560 012

MS received 14 February 1977; revised 20 June 1977

Abstract. Simple formalized rules are proposed for automatic phonetic transcription of Tamil words into Roman script. These rules are syntax-directed and require a one-symbol look-ahead facility and hence easily automated in a digital computer. Some suggestions are also put forth for the linearization of Tamil script for handling these by modern machinery.

Keywords. Automatic phonetic transcription; data representation; linearization of Tamil script; Roman script; Tamil script; transcription.

1. Introduction

It is well known that in the oldest of the Dravidian languages, Tamil, the pronunciation of the letters corresponding to the constants க், ச், ட், த், ப், ற் (called Surds) and their consonanto-vowels* vary considerably depending upon the sequence in which these letters are embedded (hence may be called phonetically context-sensitive) e.g. there is only one set of a consonanto-vowel (CV) corresponding to the four different sound-values** in the words below; these are underlined.

நகை; கட்டு; பாக்கு பொங்கல்

Na hai Ka t tu Pakhkhu Ponggal

In other words there is a many-to-one mapping from sound-values to the Surd and a one-to-many mapping from the surd to its sound value.

Accordingly, Tamil is not a strictly phonetic language (in relation to its writing system) unlike all other Indian languages. The advantage of not being strictly phonetic in this sense is rewarding, since there is an enormous reduction in the number of letters in the alphabet; the disadvantage, however, from the point of view of a non-native speaker is the difficulty in learning the actual pronunciation which varies depending upon the context.

It is further important to study whether there exist simple formalized sound-generative rules (which are context-sensitive) for the correct pronunciation (as in

*Consonanto-vowel=a letter representing a consonant-vowel (CV) combination.

**These are the sound-values met with in the author's own dialect. Other sound values may be found in other dialects. The same method would still hold with suitable modifications in figure 2. As our aim in this paper is towards an automatic approach to the problem, we confine to as few rules as possible. We believe further studies will result in improvement.