An Investigation on Unicode Standards for Tamil

Dr. M. Ponnavaikko,
Director, Tamil Virtual University

Abstract

The problems associated with the present Unicode Standard for Tamil in its use for Tamil computing have been a big concern for the Tamil computing world. The issues are under discussion for quite sometime and almost in all the Tamil Iniayam conferences since 1997. As an alternative to the present Unicode Standard for Tamil there are two different views presented by the Tamil computing professionals. One view is to include only the vowels and the pure consonants in the encoding and the other view is to include all the 247 characters and the special Tamil symbols into the encoding scheme.

Some independent investigations in 8-bit environment were made by different groups to assess the merits and demerits of these schemes. But, no scientific investigations had been made in a 16-bit environment. These issues were considered at the State and the Central Govt. levels by the concerned departments and decided that the Govt. of Tamil Nadu will investigate the schemes under demand and submit the results for a final decision in the matter. Accordingly, it was decided that the Tamil Virtual University which has been entrusted by the Government to deal with all technical aspects related to Tamil in IT will take up the responsibility of investigating the 3 schemes, namely, (i) the present Unicode Standard for Tamil, (ii) Pure Consonant and Vowel scheme and (iii) the All Character Encoding Scheme. Thus, the studies were entrusted to a software company. The studies were completed and the results were communicated to the Ministry of Information Technology, Govt. of India for necessary reference. The study results were also communicated to the workgroup-2 of INFITT. This paper presents the different aspects and the outcome of the study. The paper includes the details on the development of 16-bit encoding schemes for the three different proposals along with the development of an Editor and an interface for inputting the Unicode Characters, the investigation results on memory requirement for storing data under the three different encoding schemes, efficiency of text processing, efficiency of processing database applications and efficiency of morphological analysis. The study has clearly brought out the positive and negative aspects of the three different proposals for the Unicode standard for Tamil.