Development and implementation of Tibetan collation algorithm

Tashi Tsering
W.N. Martin

There are four fundamental issues for a language to be processed or supported in a computer system: encoding, font design, input method, and collation algorithm. Without a collation algorithm, a computer cannot collate the strings of a language, and hence the language cannot be processed in a computer thoroughly. For those languages that use Latin letters, a collation algorithm is not an issue. However, for those complex languages – languages which have a more complex writing structure – a collation algorithm could be very complicated. Some languages in the world are still not supported in major operating systems such as Microsoft Windows. Part of the reason for this is that problems surrounding the collation issue for these languages have yet to be solved, and consequently, the computer systems are unable to thoroughly support the languages. Tibetan is one such complex language. It has many of the same problems in the design and implementation of a collation algorithm possessed by some other complex languages. Although a popular opinion of the past was that it was not possible to devise a sufficiently robust algorithm for the collation of Tibetan strings in a computer, in this paper, we will present a Tibetan collation algorithm deduced and designed from traditional rules for sorting Tibetan syllables. In addition, we will discuss the implementation issues of the algorithm, Java source code for the implementation of the Tibetan algorithm already having been opened to the public

<http://www.cs.virginia.edu/~tt3e/files/Research.html>.

A demonstration of sorting Tibetan words for a sample dictionary using this algorithm will be presented as well.