

INPUT-OUTPUT ISSUES IN SOFTWARE LOCALIZATION OF SOUTH ASIAN LANGUAGES

Bidyut B Chaudhuri
Indian Statistical Institute
203 B T Road
Kolkata-700108
bbc@isical.ac.in

Input-Output to computer system in South Asian languages is an important localization issue for the expansion of IT and software activities in these regions. There are several problems to be addressed towards this end. The first problem encountered is that of keyboard design. Since South Asian languages in general, and Indian language alphabets in particular, have more number of characters, the QWERTY keyboard structure is not well-suited for data entry. Re-configuring this keyboard creates one or other kind of problem to the PC users and data entry operators. Also, in the softwares marketed by the vendors, hardly any logically justified rule of re-configuration is employed. The general pattern followed by them is to assign the key for a English character, say K, to an Asian language character that sounds nearest to K. If the phonetic unit for a character in local language has no match in English alphabet, somewhat arbitrary key assignment is made for that character. There is no uniform norm followed by the vendors, although some of them allow the user to reconfigure the keyboard according to his/her choice. Others do not do so, hence the operator accustomed to typing with a particular software finds it difficult to work with that from another vendor. Even those who allow re-configuration, do not have uniform way of generating the vowel diacritical marker or the compound character, (a speciality of many Asian language scripts). In general, it requires the use of shift key many more times in printing south Asian scripts compared to English. As a result the work becomes slow and inefficient.

To tackle the problem, some insight in the keyboard design and the word-processing software is needed. A completely new design and manufacture of keyboard is risky for the hardware company unless key stroking for English is as simple as it is in standard QWERTY keyboard. This is so because majority of computer based activities are still done in English alphabet-based language. So, continued use of QWERTY keyboard is justified for language localization. But it is necessary to form a strong co-ordinating body who can work on and then recommend a format and fix the keys for writing basic, modified and compound characters in the most convenient way. Arabic based scripts are of special and separate concern, while the problem may be simpler for Punjabi and Tamil where there is no compound character in the writing system.

In finding a uniform format stated above, sound similarity with English should not be the only concern. The physical effort needed for finger movement while key-stroking should also be optimised. This can be designed using the character level frequency (i.e. character probability measure) in respective language. The frequency can be computed from a language corpus of reasonable size, say one million words. The characters can be ranked in descending order of frequency and those on the top should be placed in the keyboard so that they are easily reachable by the forefingers.

Middle finger gets somewhat lower priority and takes care of characters of lesser frequency. Economy in finger movement can be achieved in this manner and efficiency can be increased in the long run.

The next problem to consider is the design of font and its software representation. There is an aesthetic issue combined with the technical issues in font design of any alphabetic system. Beauty and readability are main aesthetic issues for any font. While beauty cannot be defined in quantitative terms, many aspects of readability can be quantified and tested in a scientific way. There is hardly any serious quantitative study on reading speed and ease, and their dependence on the shape, and size of the character, word and line spacing as well as paper and print quality of South Asian language scripts. It is only grossly known that for easy reading, the font size should be about 2 point larger than English font size having similar comfort and the line spacing, because of the vowel modifiers and compound characters, should also be larger. Moreover, the effects of generating some characters by economizing on glyphs have not been studied well for these languages. In many popularly marketed fonts, some character combination creates touching between neighbouring characters at unauthorized points and also make the display of funny combination of glyphs possible. Such possibilities make the font aesthetically less acceptable. But more serious problem is the impossibility of rendering some compound character combination in some fonts available in the market.

Though not well researched, a reasonably large number of fonts are now marketed in many scripts of South Asian languages, most notable among them being Devanagari, Bangla and Tamil fonts. Many fonts in languages like Telugu and Kannada are also coming up. Again, there is no uniformity in the glyph set of the fonts developed by different vendors for the same language printing system. Another angle of the problem is that actual rendition of the compound characters have not been standardized in several scripts eg. Bangla. So, two different vendors' fonts may display and print the same compound in two different ways. A third problem is the internal coded representation of the text entered by the operator. A code called ISCII (Indian Standard Code for Information Interchange) was proposed to represent most of the official languages of India, but this is not true for all major scripts of South Asia. Even for Indian scripts, most font manufacturers either do not use or do not make the ISCII version of the files transparent and usable by those who want to develop language technology softwares. As a result, a large amount of documents being made regularly could not be used language for corpus study or other application oriented research. For example, the non-English Indian newspaper websites do not provide an ISCII version of their daily news.

In this connection, the recent proposal of UNICODE standard should be mentioned. The standard is yet to be finalized and implemented, but several meetings have been organized for finding the best coding set for each language scripts. One such held in New Delhi during 24-26 September 2003 on Indian scripts may be mentioned here. During the discussions it appeared that the committee did not take appropriate expert group opinion in fixing the codes. For some Indian languages several ancient and obsolete characters, especially their compound forms have occupied the coding table (current version 4.0) unnecessarily. On the other hand, different literary institutions like *Sahitya Academy* of various states as well as State Governments did not show sufficient interest in participating in the UNICODE standardization program. But such participation is important for the advancement of

localized computer usage, as well as for furthering R & D activities in these languages.

Creation of websites in local languages is an important application area that needs special attention. They not only provide information and data to the user but also useful in business and customer services as well as in e-governance. Though there are many vendors of local language fonts, most of their softwares are suitable for DTP class of applications. Creation of high quality web portals are still limited to a few portal shops. Even some big business houses and newspapers in local language put their text matters in picture format. This is a waste of bandwidth and memory. Also, the information cannot be easily used as textual data for further sophisticated computational processing.

In fact, to create web pages in Asian language, there should be a mechanism so that the fonts are available in the machine of the web user. The user can get it in three possible ways: manual download, automatic download and dynamic delivery. In the first process, the user manually downloads the font and install it before looking the site. This is a cumbersome and unattractive process. In automatic downloading approach, the user downloads a compact file which by itself gets installed in the local machine. Such mechanism may be prone to virus etc. A better mechanism is the dynamic delivery where the font information is available only when the relevant web page is being visited. There are two technologies to support dynamic delivery. One is the PFR (Portable Font Resource) based on the web font technology from Bit Stream that can be coupled with Netscape Browser and also with Internet Explorer. The other is the EOT (Embedded Open Type) from Microsoft's Web Embedding Font Technology that is suitable for Internet Explorer only. Netscape browsers do not support the EOT. Very few vendors have web enabling technology in Asian languages. In India, CDAC and Modular Infotech (both from Pune), among others, have such capabilities.

Current statistics show that websites in local languages are coming up faster than in English. Sites in south Asian languages, especially those in Indian languages are growing at a very rapid pace. The popular search engines currently available in the market are suited to search for sites in English. A strong need is now being felt to develop search engines in multiple languages including those of South Asia. Work has started at Indian Statistical Institute in this direction and to start with English, Bangla and Hindi languages are considered in our experiment. Since some sites host textual documents in picture mode, a modified version of OCR engine is also being designed to read the textual portions from such documents.

Like English and other European languages, it is important also to develop the digital library in South Asian languages. There are two ways of achieving this goal. One is the manual entry, which is hampered by inadequate Word-processors available in local languages. An example is the 3 million-word corpus created about five years ago in various major Indian languages using GIST technology, where several compound characters could not be rendered properly. Even now the situation has not substantially improved, as evident from the Hindi word-processing environment of the most popular software vendor Microsoft. The other way of quick data entry from printed matter is the use of OCR device. Recently, some limited font trainable OCR systems have been reported on Devanagari, Bangla, Tamil and Punjabi (which is a simplified version of Devanagari) scripts. OCR systems in other scripts like Oriya,

Telugu, Malayalam and Kannada are being researched in some R & D organizations. Efforts to develop OCR for Arabic based Scripts like Urdu, Kashmiri and Sindhi are also ongoing in parts of Pakistan and India. A major difficulty of developing a good OCR system in most Indian languages lies in the large set of basic, modified and compound characters in the text, numbering about 350. Also, for some scripts the characters get topologically connected in a word. So, prior character segmentation is necessary to input the data in the character-based recognition engine. Yet, for laser quality printing with some fonts, up to 98% accuracy has been achieved in Devanagari and Bangla OCR system. Indian Statistical Institute has done pioneering work in this field. The Devanagari system has been brought to the market by CDAC (who purchased the technology from ISI) in the brand name *Chitrakan*. The problem of Tamil and Punjabi OCR is more simple because of smaller number of character shapes to be recognized, and a Tamil OCR system is also available in the market. Now the work at this institute is being extended to develop a multi-lingual OCR system which would cover some of the major Indian scripts. The work involves automatic identification of various text in multi-text document and send the respective portions to their corresponding OCR modules. Some initial pre-processing steps like skew correction, binarization, line, word and character segmentation may be common for several scripts eg Bangla and Devanagari. Nevertheless, the OCR based data input procedure in South Asian languages have not yet matured, and a substantial amount of research is needed to achieve this.

Inputting linguistic data and information in computer can also be done by electronic pen and tablet devices. For this purpose, some online handwritten character recognition software is needed. To the best of our knowledge, there is only one PhD study done in Indian script online handwritten character recognition. Recently some work has started at a few places including Indian Statistical Institute, where recognition of handwritten Bangla text has been tested with some success. Again, work on handwritten Indian scripts are more difficult than on Roman-based scripts because of topological complexity of compound characters, whose ideal forms can be scripted by a few highly educated persons only. Most writers would pen the characters in their own suitable way, and the most powerful recognition engine may be deceived by such characters. This would make offline handwritten OCR system even more difficult and research effort in this direction is still at its infancy. Rather, some progress has been achieved for recognizing table-form documents where a person has to input some data by pen in the specified box and individual characters are isolated by the box boundary. People are more careful in writing within boxes, since their personal interest is involved in filling such forms. This makes the recognition problem simpler than free-form handwriting recognition. In any case, there is a need of further concentrated effort to develop such input device in South Asian languages.

The other form of linguistic data input can be done through a speech recognition system. Here the system understands what has been spoken and converts it into text version automatically. However, the most advanced speech recognition systems developed so far is poorer than all other input systems in reliability and accuracy irrespective of language. However, some research labs in the west have announced their programs on Speech-to-Speech machine translation system. We believe that there is a long way to go before any concrete success is achieved in that direction. In south Asian language context the speech recognition research did not gather sufficient momentum and most research labs are still working on isolated

spoken word recognition only. The only advanced level work worth mention is that done at IBM lab at New Delhi, where successful demonstration of continuous Hindi speech recognition has been given in controlled environment.

While speech as input device is still poor, output device based on speech synthesis is a mature technology. In south Asian countries, especially in India, speech synthesis softwares have been developed in various languages. All of them are based on simple di-phone concatenation technology with little grapheme to phoneme rules incorporated. Almost no attempt is made there to import prosody in the synthesized speech. The Indian pioneers in speech synthesis activities are Indian Statistical Institute (Bangla speech) and Tata Institute of Fundamental Research (Hindi speech). Several application softwares like talking dictionary, Text to Braille and Braille to Text with speech synthesis output have also been developed at ISI and the system has been taken by various organizations for commercial distribution.

To improve the quality of synthetic speech, exhaustive utterance rules for all popular words should be found for any language. The problem is not so simple to solve, as experienced by us for Bangla. The other alternative is to store a spoken form of each word in the dictionary and use this phonetic representation for concatenation of speech signals. Study of prosody of spoken language is another important research area, where very little has been done in south Asian languages. Also, since computer memory is cheap and processors are now faster, one can employ syllable concatenation instead of di-phones and get a more naturally audible synthetic speech. Work has started in this direction at Indian Statistical Institute.

To conclude, it may be generally observed that generation of computer output using different media like text, image, drawing and speech is easier than computer input using these media, where intelligent recognition is involved. So, more effort is needed for developing various inputting softwares. We have to remember this for localization of south Asian languages as well.