# Romanized Indic and LaTeX

Anshuman Pandey

## 1 Introduction

In 1990 at the 8th World Sanskrit Conference in Vienna, a panel of Indologists devised two encoding schemes which would enable them to exchange electronic data across a variety of platforms. These schemes are the “Classical Sanskrit” and “Classical Sanskrit eXtended” encodings, widely known in Indological circles as CS and CSX, respectively, or simply, CS/CSX.

## 2 CS and CSX

The CS and CSX encodings are currently the closest thing to an accepted standardization of the 8-bit transliteration of Indic scripts. CS/CSX is based on IBM Code Page 437, whose characters of the range 129–255 have been reassigned with characters traditionally used for the romanization of Sanskrit.

The accented French and German characters in the cp437 range 129–223 were not altered, in order to facilitate the input of these languages as well as English and Sanskrit. The accented characters required for Sanskrit were located as far as possible in the positions used by cp437 for graphic or mathematical symbols.

The re-encoding of cp437 was discussed in a document by Dominik Wujastyk titled *Standardization of Sanskrit for Electronic Data and Screen Representation* [1].

CS is a basic inventory of diacritic letters comprising the following characters traditionally used for the transliteration of Classical Sanskrit:

l̃ ṁ ā Ā ī Ī ū Ū ṛ
Ṛ ṝ Ṝ ḷ Ḷ ḹ Ḹ ṅ Ṅ
ṭ Ṭ ḍ Ḍ ṇ Ṇ ś Ś ṣ
ṃ Ṃ ḥ Ḥ

CSX is an extension of the above which provides the following additional characters used in Vedic Sanskrit and in Prakrit:

r̲ ā̆ ī̆ ū̆ n̲ ā́ ā̀ ī́ ī̀
ū́ ū̀ ṛ́ ṛ̀ ṝ́ ã ĩ ũ ē
õ ĕ ŏ l̲

Contrary to what the name indicates CS/CSX is not limited to the transliteration of Sanskrit, and may be used to transliterate many other Indic scripts effectively.

## 3 ISO 15919 and CSX+

ISO/TC46/SC2/WG12, the International Standards Organization Working Group for the Transliteration of Indic, has been busy with the draft ISO 15919 standard [2]. This draft standard provides tables which enable the romanization of Indic scripts which are specified in Rows 09–0D and 0F of UCS (ISO/IEC 10646 and Unicode).

This romanization is accomplished using plain ASCII 7-bit (ISO-646) characters, two or three roman characters often being required to represent a single Indic one. These tables provide for the Devanagari, Gujarati, Gurmukhi, Bengali (including Assamese), Oriya, Telugu, Kannada, Malayalam, Tamil, and Sinhala scripts. This draft is not yet a standard, although work is well advanced.

While ISO 15919 is still in draft stages, it appears that a consensus has been reached with regard to the form of transliteration. It was therefore decided that CS/CSX ought to be further extended to account for the new characters proposed in the draft standard. John Smith, Dominik Wujastyk, and I developed an extension to CS/CSX known as CSX+ (Classical Sanskrit eXtended+).

CSX+ aims to be downward compatible with CS/CSX, save for the relocation of two characters in positions used for non-breaking spaces in popular software packages. While seeking to implement the draft ISO 15919 standard, CSX+ also retains a useful set of European accented characters, dashes, and quotes.

Most of the new characters are those required for the draft ISO 15919 standard, which specifies the following characters which are unsupported by CS/CSX:

ǣ Ǣ ŭ r̥ R̥ ŕ̥ r̥̀ r̥̄ r̥̄́
l̥ l̥̄ ē Ē ē̃ ō Ō ō̃ ẏ
r̆ m̐ n̆ m̆ t̲ k̲ k̲h̲ K̲h̲ ġ
Ġ č Č h̲ ḫ

In addition to the above, the following further characters have been added as being centrally useful in any text encoding:

“ ” – —

There is a single “European” accented character — ÿ — that CSX inherited from the original Code Page 437, but that is unlikely to be required for any Indian or European language. It has been eliminated to save one character slot. Other characters removed from the code page are the currency symbols sterling, yen, and cent, and the guillemets.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 128 | Ç | 147 | ô | 165 | Ñ | 184 | ī̀ | 203 | Ǣ | 222 | — | 241 | ṭ |
| 129 | ü | 148 | ö | 166 | l̃ | 185 | ē | 204 | k̲h̲ | 223 | “ | 242 | Ṭ |
| 130 | é | 149 | ò | 167 | ṁ | 186 | ō | 205 | ġ | 224 | ā | 243 | ḍ |
| 131 | â | 150 | û | 168 | ā̆ | 187 | R̥ | 206 | ĉ | 225 | ß | 244 | Ḍ |
| 132 | ä | 151 | ù | 169 | ī̆ | 188 | ẏ | 207 | ṝ́ | 226 | Ā | 245 | ṇ |
| 133 | à | 152 | ǣ | 170 | ū̆ | 189 | ū́ | 208 | ã | 227 | ī | 246 | Ṇ |
| 134 | å | 153 | Ö | 171 | ā̃ | 190 | ū̀ | 209 | ĩ | 228 | Ī | 247 | ś |
| 135 | ç | 154 | Ü | 172 | ī̃ | 191 | ř | 210 | ũ | 229 | ū | 248 | Ś |
| 136 | ê | 155 | ŭ | 173 | ṉ | 192 | ō̃ | 211 | ẽ | 230 | Ū | 249 | ṣ |
| 137 | ë | 156 | ē̃ | 174 | r̥̄ | 193 | m̐ | 212 | õ | 231 | ṛ | 250 | Ṣ |
| 138 | è | 157 | r̥ | 175 | l̥ | 194 | ṯ | 213 | ĕ | 232 | Ṛ | 251 | ” |
| 139 | ï | 158 | á | 176 | l̥̄ | 195 | Ē | 214 | ŏ | 233 | ṝ | 252 | ṃ |
| 140 | î | 159 | ṟ | 177 | ŕ̥ | 196 | Ō | 215 | ḻ | 234 | Ṝ | 253 | Ṃ |
| 141 | ì | 160 | *space* | 178 | r̥̀ | 197 | ň | 216 | ū̃ | 235 | ḷ | 254 | ḥ |
| 142 | Ä | 161 | í | 179 | r̥̄́ | 198 | ṛ́ | 217 | Ġ | 236 | Ḷ | 255 | Ḥ |
| 143 | Å | 162 | ó | 180 | m̆ | 199 | ṛ̀ | 218 | Ĉ | 237 | ḹ | | |
| 144 | É | 163 | ú | 181 | ā́ | 200 | K̲h̲ | 219 | ẖ | 238 | Ḹ | | |
| 145 | æ | 164 | ñ | 182 | ā̀ | 201 | ḵ | 220 | ḫ | 239 | ṅ | | |
| 146 | Æ | | | 183 | ī́ | 202 | *space* | 221 | – | 240 | Ṅ | | |

Table 1: CSX+ Character Encoding Table

The remaining assignments have been made on the basis that the best use for the small number of spare slots available is to use them for capitalised versions of those new characters with the most need for capital forms — i.e., characters capable of beginning a word.

## 4 Input encoding

As the use of LaTeX amongst Indologists has significantly increased, I felt that the CSX+ encoding scheme ought to be adapted for use with LaTeX through the `inputenc` package. To serve this end an input encoding definition file called `cp437csx.def` has been developed and placed on CTAN in the directory `fonts/csx/styles/`. Such an input encoding definition will make the typesetting of romanized Indic much easier, and might lead to the development of hyphenation patterns for romanized Sanskrit and other Indic languages.

The file `cp437csx.def` enables text encoded in CSX+ to be read and accurately typeset by LaTeX without the need for converting the CSX+ text into LaTeX accent codes. Table 1 provides a map of the CSX+ encoding character set. A screen font and driver for displaying CSX+ text on MS-DOS and OS/2 systems is also available in the directory `fonts/csx/`.

The stabilization of the CSX+ encoding, in tandem with the emerging ISO standard, will encourage further necessary work, such as hyphenation tables for romanized Indic.

### References

[1] Wujastyk, Dominik. *Standardization of Romanized Sanskrit for Electronic Data Transfer and Screen Representation* [results of a session held at the 8th World Sanskrit Conference, Vienna, 1990], in *Sesame Bulletin* 4(1), 1991, pp. 27-29. Also available as a PostScript document from `CTAN/fonts/csx/csx-doc.ps`.

[2] Stone, Anthony [ed]. *ISO Committee Draft 15919: Transliteration of Devanagari and Related Scripts into Latin Characters.* Available at `http://ourworld.compuserve.com/homepages/stone_catend/trdcd1c.htm`.

◇ Anshuman Pandey
University of Washington
Department of Asian Languages
and Literature
225 Gowen Hall, Box 353521
Seattle, WA 98195
`apandey@u.washington.edu`
`http://weber.u.washington.edu/~apandey/`