Standardization of Sanskrit for Electronic Data Transfer and Screen Representation

Dominik Wujastyk

9 September 1990

Text Encoding Guidelines

During the 8th World Sanskrit Conference, Vienna 1990, a panel was held to discuss the standardization of Sanskrit for electronic data transfer. Participants were encouraged to acquire and study the *ACH-ACL-ALLC Guidelines for the Encoding and Interchange of Machine-readable Texts*, edited by Lou BURNARD and C. M. SPERBERG-MCQUEEN (Chicago and Oxford, 1990). These *Guidelines* are available free of charge in Europe from L. Burnard, Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, England, or in the USA from C. M. Sperberg-McQueen, Computer Center (MIC 135), University of Illinois at Chicago, Box 6998, Chicago, IL 60680, USA.

7-bit coding for file transfer

Professor H. Falk presented a program called `CONVERT` that conveniently converts any coding scheme used in a data file to any other coding scheme. This program was generously made available at no cost, together with Turbo Pascal source code. Prof. Falk also presented a very useful 7-bit, multi-byte "mediation code" which will be of general use for file exchange.

8-bit character set for text display

Finally, although the above two provisions cover all essential needs, the panel still felt that a standard assignment of graphic codes for the display of Sanskrit transliteration would be helpful. An ad hoc committee of interested parties was formed, and two 8-bit 'code pages" were designed. One, *Classical Sanskrit* (CS), for standard use and another, *Classical Sanskrit Extended* (CSX), which included the former, but also provided for Vedic, MIA, Tamil and some special usages. These code pages take as their point of departure IBM's code page 437, the default set of character codes built into the IBM PC and clones. The characters listed below are replacements for the characters in code page 437 which have the same numerical code. E.g., character number 224 in code page 437 is a Greek letter alpha (α); CS redefines it to be a with a macron (ā). All codes not specified below are assumed to be as code page 437. E.g., character number 130 is e acute (é).

The codes assigned were as follow:

Classical Sanskrit (CS)

166 l tilde	l̃	240 N overdot	Ṅ
167 m overdot	ṁ	241 t underdot	ṭ
224 a macron	ā	242 T underdot	Ṭ
225 not used (normally German *eszett*, ß)		243 d underdot	ḍ
226 A macron	Ā	244 D underdot	Ḍ
227 i macron	ī	245 n underdot	ṇ
228 I macron	Ī	246 N underdot	Ṇ
229 u macron	ū	247 s acute	ś
230 U macron	Ū	248 S acute	Ś
231 r underdot	ṛ	249 s underdot	ṣ
232 R underdot	Ṛ	250 S underdot	Ṣ
233 r underdot macron	ṝ	251 not used (normally the root sign $\surd$)	
234 R underdot macron	Ṝ	252 m underdot	ṃ
235 l underdot	ḷ	253 M underdot	Ṃ
236 L underdot	Ḷ	254 h underdot	ḥ
237 l underdot macron	ḹ	255 H underdot	Ḥ
238 L underdot macron	Ḹ		
239 n overdot	ṅ		

Classical Sanskrit Extended (CSX) additions

The following definitions are added to the above Classical Sanskrit character set.

159 r underbar	r̲	199 r underdot grave	ṛ̀
168 a macron breve	ā̆	207 r underdot macron acute	ṝ́
169 i macron breve	ī̆	208 a tilde	ã
170 u macron breve	ū̆	209 i tilde	ĩ
173 n underbar	n̲	210 u tilde	ũ
181 a macron acute	ā́	211 e tilde	ẽ
182 a macron grave	ā̀	212 o tilde	õ
183 i macron acute	ī́	213 e breve	ĕ
184 i macron grave	ī̀	214 o breve	ŏ
189 u macron acute	ū́	215 l underbar	l̲
190 u macron grave	ū̀		
198 r underdot acute	ṛ́		

These codes were chosen to have minimal impact on the standard IBM PC extended ASCII character set, but they are intended for general use in displaying Indological texts on any machine with an 8-bit (or greater) character set.

Dr. D. Wujastyk will be making available small programs that load the above character sets into the EGA or VGA display adaptors, for IBM PC users.

The above character codings have been approved by R. E. Emmerick, H. Falk, R. Lariviere, G. J. Meulenbeld, H. Nakatani, M. Tokunaga, D. Wujastyk, P. Schreiner and M. Yano.

These character codings are primarily intended for use in situations when the screen display of these characters is requried, such as in word processing. They may, of course, be used for data transfer, where, however, a 7-bit code (perhaps with multi-byte character codes) is still preferable. One such 7-bit scheme is provided hy H. Falk (see 2. above).

These character codings are currently open for discussion and comments may be directed to Dr. D. Wujastyk at

Wellcome Institute,
183 Euston Road,
London NW1 2BN, England,

or by email at

Bitnet/Earn: `dow@harvunxw` or
Janet: `D.Wujastyk@uk.ac.ucl`.

After a suitable lapse of time, the character sets will be sent to ECMA and ISO for registration. They will also be sent to the Text Encoding Initiative for registration, probably with H. Falk's 7-bit coding scheme.

Such registration in no way enforces these schemes; it merely makes them available centrally for reference. Other schemes may also be registered in the future.