Indica, an Indic preprocessor for TeX. A Sinhalese TeX System

Yannis Haralambous

Indica, an Indic preprocessor for TEX
A Sinhalese TEX System

Yannis Haralambous

Abstract

In this paper a two-fold project is described: the first part is a generalized preprocessor for Indic scripts (scripts of languages currently spoken in India—except Urdu—, Sanskrit and Tibetan), with several kinds of input (LATEX commands, 7-bit ASCII, CSX, ISO/IEC 10646/UNICODE) and TEX output. This utility is written in standard Flex (the GNU version of Lex), and hence can be painlessly compiled on any platform. The same input methods are used for all Indic languages, so that the user does not need to memorize different conventions and commands for each one of them. Moreover, the switch from one language to another can be done by use of user-defineable preprocessor directives.

The second part is a complete TEX typesetting system for Sinhalese. The design of the fonts is described, and METAFONT-related features, such as metaness and optical correction, are discussed.

At the end of the paper, the reader can find tables showing the different input methods for the four Indic scripts currently implemented in *Indica*: Devanagari, Tamil, Malayalam, Sinhalese. The author hopes to complete the implementation of Indic languages into *Indica* soon; the results will appear in a forthcoming paper.

– – ∗ – –

1 *Indica*

1.1 Introduction

Many Latin-alphabet native writers find the Greek and Cyrillic alphabets exotic (not to mention African and phonetic characters). Actually this shouldn't happen, since —at least for the upper case— Greek, Cyrillic and Latin types can have the same design: they have the same roots, have evolved more-or-less in the same way, and the same principles of Occidental type design can be applied to them. There are even common glyphs to the three ('A', 'B', 'E', 'H', 'M', 'O', 'P', 'T', 'X') which will appear only once in case one wishes to have a big "Greco-Cyrillico-Latin" font.

The situation is completely different in the case of Indic languages. Once again all of their scripts have the same roots, but instead of keeping the same style and being complementary to each other, they all have the same set of letters, in the same order, but with (often very) different shapes. Every child in India learns the same alphabet "ka-kha-ga-gha-..." but depending on the region, the letter shapes can be very different: क ख ग घ ङ च छ... in Devanagari script, ಕ ಖ ಗ ಘ ಙ ಚ ಛ... in Malayalam, ක ඛ ග ඝ ඞ ච ඡ... in Sinhalese, etc.[1]

This justifies the choice of a common transliteration scheme for all Indic languages. But why is a preprocessor necessary, after all?

A common characteristic of Indic languages is the fact that the short vowel 'a' is inherent to consonants. Vowels are written by adding diacritical marks (or smaller characters) to consonants. The beauty (and complexity) of these scripts comes from the fact that one needs a special way to denote the *absence of vowel*. There is a notorious diacritic, called "virāma", present in all Indic languages, which is used for this reason. But it seems illogical to *add* a sign, to specify the *absence* of a sound. On the contrary, it seems much more logical to remove something, and what is done usually is that letters are either brought very near (in Sinhalese) or written one over another (Malayalam), or written together while losing some parts (Devanagari, Bengali, ...). In this way we obtain those hundreds of beautiful ligatures which make the charm of Indic scripts.

When typesetting with TEX, the preprocessor will have to indicate to TEX all the necessary ligatures which can be either constructed from character parts (as in the case of Velthuis's Devanagari), or spread in several 256-character tables (as in the case of the Sinhalese font described in the second part of this paper). Also, it often happens that a vowel is written in front of a group of consonants, although phonologically it comes after the group; and since the transliteration is always phonetic, the preprocessor will take the vowel from where it belongs phonetically and place it where it belongs graphically.

Finally the preprocessor is needed for the simple task of inserting \- commands (discretionary hyphens) at the appropriate locations: since characters and ligatures are often constructed from other characters, or belong to several font tables, there is little hope for getting efficient hyphenation patterns so that TEX can hyphenate as it does for Western languages.

1.2 The interna of *Indica*

The preprocessor *Indica* is written in a special way, allowing easy changes and expansions, thanks to the use of **Flex**. Flex is a lexical analyzer, released under GNU copyleft; it generates C code out of simple pattern matching instructions. The advantage of

[1] One could compare this situation to the existence of Antiqua, old German, and Irish types for the same alphabet (a differentiation sadly missing from the ISO/IEC 10646/UNICODE encoding).

Flex is that without being a good programmer one can make powerful and error-free C programs.

How does it work? The minimal Flex file is of the form

```
%{
%}
%%
...lines of code...
%%
main()
{
yylex();
}
```

where the *lines of code* are of the form

```
xyz   { do_this(); do_that(); }
```

`xyz` is a pattern which may appear in the input file, and `do_this();`, `do_that();` are arbitrary C commands, executed whenever the pattern is matched in the input file. This scheme is extremely powerful, since patterns can be arbitrary regular expressions. Suppose, for example, that you want to write a program which finds all TEX commands followed by a blank and adds an empty group to them, if needed (to avoid getting TEXis beautiful, as most TEX users did at least once in their lives): `\TeX␣` shall be replaced by `\TeX{}` and so on, for every command followed by a blank. You can with the following single line of Flex code:

```
"\\"[a-zA-Z]+/" " { ECHO; printf("{}"); }
```

The double quotes indicate verbatim mode, the double backslash is the usual C notation to obtain a backslash in a string, `[a-zA-Z]+` is a regular expression meaning "one or more lowercase/uppercase letters" and finally `/" "` means "this pattern should be matched only if followed by `" "` (a blank)". The `ECHO;` command transmits the input pattern to the output, and `printf{}` adds the `{}` string.

The reader may now have realized the power and ease of use of Flex. Moreover, the generated C code is automatically optimized for the platform on which Flex is run so that one can be sure that the code will compile without problems into a quick and smooth executable.

Indica is written in Flex. To obtain an executable, you will have to run Flex first and then C. The necessary steps are explained in section 1.3.1. Having read the excellent book *lex & yacc* by Levine, Mason and Brown (1992) the user will be able to adapt *Indica* to his/her personal needs, if these are not already covered by the broad range of *Indica*'s input encodings.

1.3 Guidelines for the use of *Indica*

1.3.1 How to install *Indica*

Indica is written in Flex, the GNU version of the standard UNIX utility Lex.[2] On the server you will find executables for Macintosh and MS-DOS. If you are on some other platform, or if you want to make changes to the `indica.lex` file, you will have to compile it again. This operation consists of the following (relatively straightforward) steps:

1. run Flex on `indica.lex`, with the `-8` option:

   ```
   flex -8 indica.lex
   ```

2. Flex will create the file `lex.yy.c` (`LEX_YY.C` on MS-DOS); this is a machine-generated, C++ compatible, ANSI C code file. Run your favourite C-compiler on it, and link the result with the standard ANSI C libraries.

After having fetched or compiled your own executable of *Indica*, you can use it. For this you must prepare your document using the syntax explained in section 1.5, and run *Indica* to produce a regular TEX or LATEX file. *Indica* uses the standard C input and output streams, so you have to type `<` and `>` to redirect these streams to your files:

```
Indica < foo.inp > foo.tex
```

where `foo.inp` is the document you prepared and `foo.tex` is the TEX file *Indica* will create for you.

In this way *Indica* can be used as a filter for piping operations: if your operating system allows piping and your TEX implementation uses the standard input stream, you can systematically write `Indica < foo.inp | TeX` to pre-process `foo.inp` and run TEX on the result, avoiding thereby the creation of an intermediate TEX file.

1.4 *Indica* input schemes

TEX can handle only 8-bit fonts (fonts with 256 characters at most). This seems more or less sufficient for the needs of a certain number of Western European languages, but is definitely unsuitable for Oriental scripts like the Sinhalese one.[3] Hence, the use of a preprocessor is unavoidable. *Indica* will allow the use of the same input scheme(s) for all Indic languages: one will be able to write multilingual Indic documents without changing the input conventions, whenever a language switch occurs. There are

[2] Actually it uses a very important feature of Flex which is not part of the POSIX Lex standard, namely *exclusive states*. *Indica* has to be compiled on a Lex version with this feature; see Levine, Mason, and Brown (1992) for more details.

[3] The TEX extension Ω (Plaice, 1994; Haralambous and Plaice, 1994) will solve these (and many more) problems by using internally the UNICODE encoding, and 16-bit virtual fonts for the output.

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
0.	Usual 7-bit ASCII (ISO 646)															
1.																
2.																
3.																
4.																
5.																
6.																
7.																
8.	■	■	■	■	■	■	ä*	■	■	■	■	■	■	■	Ä*	■
9.	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	r̲
A.	■	■	■	■	ñ*	Ñ*	Ĩ	ṁ	ă	ĭ	ŭ	■	■	n̲	■	■
B.	■	■	■	■	■	á	à	í	ì	ĕ*	Ĕ*	ŏ*	Ŏ*	ú	ù	■
C.	ē*	Ē*	ō*	Ō*	ā*	Ā*	ŕ	ṛ	ň*	Ň*	ŏ̆	M̆*	■	■	■	f̲
D.	ã	ĩ	ũ	ẽ	õ	ĕ	ŏ	l̲	■	■	■	Ẽ*	Õ*	■	■	L̲*
E.	ā	ß	Ā	ī	Ī	ū	Ū	ṛ	Ṛ	ī	Ṛ	ḷ	Ḷ	Ī	Ḷ	ṅ
F.	Ṅ	ṭ	Ṭ	ḍ	Ḍ	ṇ	Ṇ	ś	Ś	ṣ	Ṣ	√	ṃ	Ṃ	ḥ	Ḥ

Code positions followed by * are extensions of CSX proposed hereby by the author. The gray square ■ denotes positions which have not yet been determined.
The author would like to point out that even if certain characters are usually not used in uppercase form, they could very well appear inside all-caps text; so, IHHO, all characters should be included in the table in lowercase and uppercase form. Uppercase letters missing from the table are: R̲, Ĩ, Ṁ, Ă, Ĭ, Ŭ, N̲, Á, À, Í, Ì, Ú, Ù, Ŗ, Ṛ, R̄, Ã, Ĩ, Ũ, Ĕ and Ŏ (a total of 21 codes).

Table 1: The CSX 8-bit input encoding

four possible input schemes, common to Hindi, Sanskrit, Bengali, Tamil, Telugu, Malayalam, Kannada, Oriya, Gujarati, Gurmukhi, Sinhalese and Tibetan:

1. `SEVENBIT`, a 7-bit (ISO 646) encoding scheme, based on Frans Velthuis' Hindi/Sanskrit transcription. Some extensions were necessary for Sinhalese, but also for other Indic languages, to obtain the character set of the (Indic part) of UNICODE/ISO 10646-1 standard (ISO, 1993).
2. `CSX`, the *Classical Sanskrit Extended* encoding, an 8-bit extension of ISO 646, proposed by an ad hoc committee, at the 8th World Sanskrit Conference, in Vienna 1990 (Wujastyk, 1991) (Table 1). For Sinhalese and other Indic languages some necesssary extensions were included in the character set of the (Indic part) of UNICODE/ISO-IEC 10646-1 standard (ISO, 1993).
3. `LATEX`, a standardized form of LaTeX commands (for example, only `\d{m}` is valid for 'ṃ', and not `\d m` or `\d{}m` or `\def\foo{\d{m}}\foo`, etc.), describing the "standard" transliteration of Indic languages.
4. `UNICODE`, the 16-bit version of ISO/IEC 10646-1 (see ISO, 1993), with an anticipated Sinhalese encoding by the author (since Sinhalese is not yet part of ISO 10646).[4]

[4] Although there is not a broad choice of UNICODE-compatible software yet (Windows NT is the most popular case of such software), the author believes that UNICODE is already now the ideal solution for *document storage* and *transmission*, especially when used in conjunction with a markup language like SGML.

The reader will find a complete table of equivalences between (1), (2) and (3), applied to Sinhalese, in Table 4.

1.5 The *Indica* syntax

Three kinds of predefined *Indica* commands exist:

1. commands affecting the input mode:
   ```
   #SEVENBIT
   #CSX
   #LATEX
   #UNICODE
   ```
   as described in 1.4.
2. commands determining the current (Indic) language:
   ```
   #BENGALI
   #GUJARATI
   #GURMUKHI
   #HINDI
   #KANNADA
   #MALAYALAM
   #ORIYA
   #SANSKRIT
   #SINHALESE
   #TAMIL
   #TELUGU
   #TIBETAN
   #NIL
   ```
   the last one being used to return for arbitrary non-Indic text to non-preprocessed mode.
3. the
   ```
   #ALIAS
   ```
   command, which allows creation of new names for the commands listed above.

Here are the rules you have to follow when using these commands:

- the "escape character" for *Indica* commands (or should I say "directives"?) is `#`. A command name consists of this character, followed by at most 32 *uppercase letters* or *8-bit characters* (in the range `0x80`–`0xff`). It follows that you can write, for example, '`#NIL;`' or '`#NILthis`', but *not* '`#NILYannis`'; in the latter case you can either leave a blank space ('`#NIL␣Yannis`') or insert an empty group ('`#NIL{}Yannis`') or apply any other similar TEXtrick.
- TEX and LATEX commands are not affected by the preprocessor. Be careful, though, because command *arguments* will nevertheless be preprocessed: if you write

  ```
  #HINDI mohan \TeX\ raake"s
  \begin{center} mis paal
  ```

  then, `\TeX` and `\begin` will be left unchanged by the preprocessor, while `center` will produce चेनतेर and `\begin{चेनतेर}` is hardly something standard LATEX would accept. In these cases it is advised to write

  ```
  #HINDI mohan \TeX\ raake"s
  #NIL\begin{center}#HINDI mis paal
  ```

- *Indica* commands are *not* nested: if you switch to Bengali and then Hindi, you will have to type `#BENGALI` once again to return to the former language (there is no "group closing" command, bringing you back to the state you were before, as in TEX for example).
- Input mode switching commands (`#SEVENBIT`, `#CSX`, etc.) can appear anywhere in the text. They don't produce any immediate effect when in `NIL` language; the corresponding input mode is stored and applied on forthcoming Indic text. Default settings (applied automatically at the begining of every file) are the `NIL` language, and `SEVENBIT` input mode.
- The `ALIAS` command has the following syntax:

  ```
  #ALIAS SINHALESE FOO
  ```

  which has to be written *at the beginning of a line*. The first argument is the command name for which we want to create an alias; the second argument is the alias itself. After the definition above, you can use `#FOO` instead of `#SINHALESE`.

  You can use *uppercase* Latin alphabet letters, or *8-bit characters* in aliases. For example, you could define

  ```
  #ALIAS MALAYALAM M
  #ALIAS NIL N
  ```

  and afterwards type only `#M` to switch to Malayalam, and `#N` to switch back to `NIL` language. Or, you could define

  ```
  #ALIAS MALAYALAM മലയാളം
  ```

  provided your platform has a graphic interface allowing Sinhalese screen display (Macintosh, Windows, X-Window...) and provided the encoding you use places Malayalam characters in the upper 8-bit range.

  Numbers cannot be part of aliases, so the usual TEX operators `#1`, `#2`, `##1`... are not affected by *Indica*. More generally, whenever *Indica* encounters a hash mark followed by an unknown string (not a predefined command name or previously defined alias), it leaves both the hash mark and the string untouched.[5]
- *Indica* does not take TEX comment marks into consideration. If you write

  ```
  % This is a TeX comment
  %#TIBETAN
  % etc etc
  ```

  unlike TEX, *Indica* will read these lines and switch to Tibetan language.
- *Indica* will read only the files you ask it to read; it will not interpret (LA)TEX `\input` commands.[6] On the other hand, a file already processed by *Indica* does not contain any *Indica* commands any more, so that you can re-process it an arbitrary number of times without altering it. It follows that you could write a batch file to run *Indica* on all files of your working directory, just to be sure that no file has been left unprocessed.

1.6 Simultaneous text and transcription

If you write your Sinhalese text in `LATEX` input mode, you can copy and paste it to some other part of the document and run it in `NIL` language mode; it will produce the "standard" Latin transcription of the same text. The only precaution you need to take is to include Christina Thiele's TEX macro `\diatop` (see Thiele, 1987), in the preamble of your document. This macro typesets characters with double or triple diacritization (like $\bar{\bar{a}}$, $\bar{\dot{r}}$, etc.)

[5] !'Cuidado! If you mispell an *Indica* command, you will end up with a hash mark and the misspelled string in your (LA)TEX code and should prepare yourself to get a very mean (LA)TEX error message: (LA)TEX just hates useless hash marks.

[6] This feature could be implemented in *Indica*, but would result in a loss of portability: every TEX implementation has its own environment variables for file path searching. The same environment variables should be included into *Indica*'s code, so that exactly the same files may be found and opened.

Here is the expansion of this macro:

```
\def\diatop[#1|#2]{{\setbox1=\hbox{{#1{ }}}%
        \setbox2=\hbox{{#2{ }}}%
        \dimen0=\ifdim\wd1>\wd2\wd1\else\wd2\fi%
        \dimen1=\ht2\advance\dimen1by-1ex%
        \setbox1=\hbox to1\dimen0{\hss#1\hss}%
        \rlap{\raise1\dimen1\box1}%
        \hbox to1\dimen0{\hss#2\hss}}}%
%e.g. of use:
%  \diatop[\'|{\=o}] gives o macron acute
```

An example of simultaneous text and transcription (illustrating the use of aliases as well) is:

```
#LATEX
#S s\d{r}tuva #N (s\d{r}tuva) season,
#S aitih\={a}sika #N (aitih\={a}sika)
historical, #S au\d{s}adha #N (au\d{s}adha)
medicines, #S \d{n}aya #N (\d{n}aya) loan,
#S ko\b{l}a\u{m}ba #N (ko\b{l}a\u{m}ba)
Colombo, \ldots
```

and (after preprocessing by *Indica*) TEX will typeset

ඍතුව (sṛtuva) season, ඓතිහාසික (aitihāsika) historical, ඖෂධ (auṣadha) medicines, ණය (ṇaya) loan, කෝළඹ (koḻaṁba) Colombo, ...

2 Sinhalese TEX

2.1 Introduction to the Sinhalese writing system

Sinhalese is one of the two major languages spoken in Sri Lanka (Ceylon), the second being Tamil. Sinhalese script is a South-Indian script, close to Malayalam and Kannada. The alphabet consists of 18 vowels and 35 consonants. It is a syllabic script: the basic consonant glyph form denotes the consonant followed by the (inherent) short vowel 'a': ක is 'ka', ඛ is 'kha' etc. To obtain a consonant not followed by a vowel, one uses a special diacritic called *virāma*. Hence, ක් is 'k', ඛ් is 'kh', etc. In contrast to Hindi, a virāma is used in all circumstances, *even* at the end of a word.

Here are the 35 Sinhalese consonants (with inherent short 'a' vowel):

ක ඛ ග ඝ ඩ ච ඡ ජ ඣ ඤ ට ඨ
ඩ ඪ ණ ත ථ ද ධ න ප ඵ බ භ
ම ය ර ල ළ ව ශ ෂ ස හ ෆ

There are also 6 nasalized consonants:

ඟ ඦ ඬ ඳ ඹ ඥ

The vowels have full-size glyphs when they appear at word beginning:

අ ආ ඇ ඈ ඉ ඊ උ ඌ ඍ ඎ
ඏ ඐ එ ඒ ඓ ඔ ඕ ඖ

A vowel following a consonant (or a series of consonants) is denoted by a special stroke, or certain auxiliary characters appearing on the right or on the left of the consonant. Here are the letters ක (ka) and ද combined with each one of these vowels:

ක කා කැ කෑ කි කී කු කූ කෘ
කෲ කෙ කේ කෛ
කො කෝ කෞ
ද දා දැ දෑ දි දී දු දූ දෘ දෲ දෙ දේ
දෛ දො දෝ දෞ

Special care must be taken in three cases:

1. When a consonant without vowel is followed by ර (r): the virāma sign of the consonant and the 'r' character are replaced by a special stroke under the consonant. For example, instead of ක්‌ර we will write ක්‍ර.

   If the *consonant* + *'r'* combination is followed by a vowel, then special rules apply. Here are the letters ක and ද (as above) combined with ර and each one of the vowels above:

   ක්‍ර ක්‍රා ක්‍රැ ක්‍රෑ ක්‍රි ක්‍රී ක්‍රෘ ක්‍රෲ
   ක්‍රෙ ක්‍රේ ක්‍රෛ ක්‍රො ක්‍රෝ ක්‍රෞ
   ද්‍ර ද්‍රා ද්‍රැ ද්‍රෑ ද්‍රි ද්‍රී ද්‍රෘ ද්‍රෲ ද්‍රෙ ද්‍රේ
   ද්‍රෛ ද්‍රො ද්‍රෝ ද්‍රෞ.

2. When a consonant without vowel is followed by ය (y): the virāma sign of the consonant disappears, and ය is replaced by the pseudo-letter ්‍ය. For example, instead of ක්‌ය we will write ක්‍ය.

   If the *consonant* + *'y'* combination is followed by a vowel, then special rules apply. Here are the letters ක and ද (as above) combined with ය and each one of the vowels above:

   ක්‍ය ක්‍යා ක්‍යැ ක්‍යෑ ක්‍යි ක්‍යී
   ක්‍යු ක්‍යූ ක්‍යෘ ක්‍යෲ ක්‍යෙ ක්‍යේ
   ක්‍යෛ ක්‍යො ක්‍යෝ ක්‍යෞ

ශා කෙෂ් කෙෂ

ද්‍ය ද්‍යා ද්‍යෑ ද්‍යෑ දුයි දුයී දුයූ දුයූ ද්‍යෘ

ද්‍යෲ ද්‍යෟ ද්‍යෳ ෙදය ෙදයි ෛදය

ෙදය්‍යා ෙදය්‍ය් ෙදය්‍යෟ.

3. A third special case occurs, when a consonant [except 'r' itself??] with inherent short 'a' vowel is *preceded* by 'r'. In that case the 'r' is not written and a spiral-like stroke is added on top of the consonant. For example, instead of ර්ල we will write ල්. This phenomenon does not occur when the consonant is followed by some other vowel than 'a'. Here are all consonants with 'r' spiral strokes:

ක් ඛ් ග් ඝ් ඞ් ච් ඡ් ජ් ඣ් ඤ් ට්

ඨ් ඩ් ඪ් ණ් ත් ථ් ද් ධ් න් ප් ඵ්

බ් භ් ම් ය් ල් ළ් ව් ශ් ෂ් ස් හ්

ෆ්

Beside the special cases enumerated above, frequently ligatures occur between consonants. A ligature between two consonants implies that the first one is not followed by a vowel; the virāma sign is omitted in that case. Here are some examples:

ඥ් + ඤ = ඥ ක් + ව = ක්‍ව

ක් + ෂ = ක්‍ෂ න් + ධ = න්ධ

න් + ථ = න්ථ න් + ද = න්ද

ත් + ථ = ත්ථ

Finally there are two special signs: *anusvara* (ṃ) written ◌ං and *visarga* (ḥ) written ◌ඃ. Sinhalese punctuation follows the English rules. Hyphenation is done between syllables, i.e., *after a vowel*.

2.2 Design of the Sinhalese font

Because of the nature of Sinhalese syllables, most combinations of consonants and vowels had to be drawn separately (the reader can find a complete table of consonant/vowel combinations in Table 3). This brought the total number of distinct character positions to more than 460, placed in three 256-character tables. Despite the large number of characters, the design of a Sinhalese font does not require a superhuman effort; in fact, the shapes of many Sinhalese letters are *modular*, and can be produced by assembling elementary strokes in different ways.

To illustrate this feature of Sinhalese letters, here is a selection of such elementary strokes[7]:

1. on the *left* side of the letter: (α) the left stem of ඝ, (β) same as α, but with an horizontal bar, as in ඩ, (γ) the left stem of ය, (δ) a lowered closed loop, as in ඹ;
2. the *middle* part of the letter: (κ) a simple baseline stroke, as in ප, (λ) the same with a pinch, as in ඩ, (μ) the same with a "bridge" as in ඝ;
3. on the *right* part of the letter: (χ) a short stroke with a rounded loop, as in ඝ, (ψ) a somewhat higher stroke with a triangular loop, as in ඵ, (ω) a high and round stroke without loop, as in ච.

Out of the combinations of these four left parts, three middle parts and three right parts we will make a table to see how many of them actually exist (NE = "does not exist"):

	α			β			γ			δ		
	κ	λ	μ	κ	λ	μ	κ	λ	μ	κ	λ	μ
χ	ප	ඝ	NE	NE	ඝ	ඝ	NE	ය	NE	NE	NE	NE
ψ	ඵ	NE	NE	ඵ	ඝ	NE	ථ	ය	NE	NE	NE	NE
ω	ච	ඩ?	NE	ච	ඩ	NE	ට	ධ	NE	ම	ඹ	ඹ

As we see, more than half of the entries represent extant characters. Similar phenomena occur for other groups of Sinhalese letters. And of course there are also some isolated cases, which have to be drawn separately (like ණ, ඔ, ළ and so forth).

This modularity of Sinhalese forms makes the choice of METAFONT for the realization of a Sinhalese font even more interesting. The Sinhalese font, as presented in this paper, was commissioned from the author by the Wellcome Institute for the History of Medicine, following a proposal by Dominik Wujastyk (to whom the author would like to express his gratitude).[8] The character forms were inspired by the font of Godakumbura (1980), compared to the forms of Disanayaka (a modern Sinhalese script method; 1993), Clough (a classical 19th century dictionary with many ligatures, 1892) and Белькович (the Russian "official" Sinhalese dictionary, 1983), the last one having the most beautiful type, in the author's humble (and non-Sinhalese native) opinion. Useful information was also found in Lambert (1983), a study of south Indian scripts, and the catalogues of writing systems of the world (Nakanishi, 1980 and Faulman, 1880).

[7] Unfortunately the author does not know the original names of these strokes.

[8] See Somadasa (1994) for the first book printed using this Sinhalese system.

	8 pt		9 pt		10 pt		12 pt	
FX	.369 pt	+6.25%	.401 pt	+2.777%	.434 pt	0%	.510 pt	−2.08%
FY	.347 pt	0%	.391 pt	0%	.434 pt	0%	.521 pt	0%
shthin	.217 pt	+12.21%	.217 pt	+10.96%	.217 pt	0%	.217 pt	−15.79%
shfat	.906 pt	+10%	.972 pt	+6.66%	.998 pt	0%	1.106 pt	−6.67%
usual_left	.406 pt	+10%	.422 pt	+5%	.434 pt	0%	.495 pt	−5%
usual_right	.406 pt	+10%	.422 pt	+5%	.434 pt	0%	.495 pt	−5%

Table 2: Scaling of font parameters for optical correction

2.2.1 Optical scaling

As we all know, one of the big advantages of METAFONT drawn characters is optical scaling, that is scaling of characters in a non-linear way, to correct certain optical effects. This technique has been applied by D.E. Knuth, in the Computer Modern fonts, the first realistic example of a font family drawn in METAFONT.

The same technique has been used for Sinhalese. Here are the (technical) details: Sinhalese characters have been designed using 6 main parameters:

1. `FX`, horizontal basic unit;
2. `FY`, vertical basic unit; (in the Computer Modern fonts the same basic unit is used horizontally and vertically, namely `u`). In cases where a length/width had to be defined independently of its orientation, we have used `.5[FX,FY]` (the mean value).
3. `shthin`, the width of thin strokes;
4. `shfat`, the width of a certain number of fat strokes; (in fact, for intermediate cases the variable quantity λ`[shthin,shfat]`, with $\lambda \in [0,1]$ has been used).
5. `usual_left`, the standard left sidebearing;
6. `usual_right`, the standard right sidebearing.

Optical correction consisted in scaling these parameters differently for 8, 9 and 12 points, as in table 2 (the reader can see in the second column the percentage of deviation from the hypothetical linearly scaled value).

As the reader can see, the value of `shthin` remains the same from 8 to 12 points; this guarantees that thin strokes will not disappear in small pointsizes (and makes letters look more elegant in large pointsizes, as in Roman Bodoni fonts). The horizontal basic unit `FX` gets (proportionally) bigger in small sizes: letters become up to 6.25% wider; `FX` also gets slightly smaller at 12 points: letters become 2.08% narrower. The same tactic is applied to sidebearings.

The following sample of text illustrates optical correction. The same text (taken from Белькович, 1983), is typeset in 8, 9, 10 and 12 point sizes.

රුසියන් සිංහල ශබ්ද කොෂය සකස්කිරීමෙහි ලා දෙස්තර දූදිගම වි. රුදිගු ගේන් ලැබුණු විශාල සහාය ගැන ඔහුට සම්පාදක වරයා කෘතඥතාවය පුද කරයි.

ශ්‍රී ලංකාව රුසියන් බස හදරන සිංහල ජනතාවටත්, සෝවියත් සංගමයෙහි සිංහල බස හදරන රුසියන් ජනතාවටත් මෙම ශබ්ද කොෂය ප්‍රයෝජනවත් වෙතියි සම්පාදක වරයා ප්‍රාර්ථනා කරයි.

රුසියන් සිංහල ශබ්ද කොෂය සකස්කිරීමෙහි ලා දෙස්තර දූදිගම වි. රුදිගු ගේන් ලැබුණු විශාල සහාය ගැන ඔහුට සම්පාදක වරයා කෘතඥතාවය පුද කරයි.

ශ්‍රී ලංකාව රුසියන් බස හදරන සිංහල ජනතාවටත්, සෝවියත් සංගමයෙහි සිංහල බස හදරන රුසියන් ජනතාවටත් මෙම ශබ්ද කොෂය ප්‍රයෝජනවත් වෙතියි සම්පාදක වරයා ප්‍රාර්ථනා කරයි.

රුසියන් සිංහල ශබ්ද කොෂය සකස්කිරීමෙහි ලා දෙස්තර දූදිගම වි. රුදිගු ගේන් ලැබුණු විශාල සහාය ගැන ඔහුට සම්පාදක වරයා කෘතඥතාවය පුද කරයි.

ශ්‍රී ලංකාව රුසියන් බස හදරන සිංහල ජනතාවටත්, සෝවියත් සංගමයෙහි සිංහල බස හදරන රුසියන් ජනතාවටත් මෙම ශබ්ද කොෂය ප්‍රයෝජනවත් වෙතියි සම්පාදක වරයා ප්‍රාර්ථනා කරයි.

රුසියන් සිංහල ශබ්ද කොෂය සකස්කිරීමෙහි ලා දෙස්තර දූදිගම වි. රුදිගු ගේන් ලැබුණු විශාල සහාය ගැන ඔහුට සම්පාදක වරයා කෘතඥතාවය පුද කරයි.

ශ්‍රී ලංකාව රුසියන් බස හදරන සිංහල ජනතාවටත්, සෝවියත් සංගමයෙහි සිංහල බස හදරන රුසියන් ජනතාවටත් මෙම ශබ්ද කොෂය ප්‍රයෝජනවත් වෙතියි සම්පාදක වරයා ප්‍රාර්ථනා කරයි.

2.3 "Do I need BigTEX for all those macros?"

Sorry to disappoint you, but there are no macros. *Indica* does all the work for you and its output is rather unreadable for a human—but quite readable for TEX. With LATEX 2ε and the `T1` (Cork) encoding you only need to place the files `T1sinha.fd`,

`T1sinhb.fd`, `T1sinhc.fd` in the same place as your other FD files, and write

```
\newcommand{\SHa}{\fontfamily{sinha}%
                  \selectfont}
\newcommand{\SHb}{\fontfamily{sinhb}%
                  \selectfont}
\newcommand{\SHc}{\fontfamily{sinhc}%
                  \selectfont}
```

in the preamble of your file. If you wish to install the Sinhalese fonts in a more formal manner, recognizing the encoding of the font as being different from `T1` (we call it `SH1`), then you only need to place files `SH1sinha.fd`, `SH1sinhb.fd`, `SH1sinhc.fd` together with the other FD files you use, and use the package `sinhala.sty` when you run LaTeX 2ε. So you would begin your document like this:

```
\documentclass{article}
\usepackage{sinhala}
\begin{document}
...
```

This method is not recommended, however, if you switch frequently from Latin to Sinhalese and your machine is not very powerful: LaTeX 2ε reads a file (called `nfsh1.def`) everytime you switch encodings; even if this file is very short, the open/close operations may slow down TeX. The author hopes that this problem will be solved in future releases of LaTeX 2ε.

If you are not working with LaTeX 2ε then you have to define the fonts manually, remembering that they always come in triplets, like

```
\font\SHa=sinha10
\font\SHb=sinhb10
\font\SHc=sinhc10
```

The available point sizes are 8, 9, 10 and 12. Please contact the author if you need other point sizes, or scale the ones you have linearly. There is no bold or slanted style yet (although it would be straightforward to obtain them out of the METAFONT code), because the author has never seen such forms. Any information on Sinhalese typographical traditions and aesthetics would be most welcome.

References

А.А. Белькович රුසියානුසිංහල ශබ්දකෝෂය *(Русско-Сингальский Словарь)*. Русский Язык, Москва, Россиа, 1983.

Rev. B. Clough. සිංහල ඉංග්‍රීසි අකාරාදිය *(Sinhalese-English Dictionary)*. Wesleyan Mission Press, Kollupitiya, Sri Lanka, 1892, facsimile edition by Asian Educational Services, New Delhi, 1982.

J.B. Disanayaka. *Let's read and write Sinhala*. Pioneer Lanka Publications, London, 1993.

C. Faulman. *Das Buch der Schrift, enthaltend die Schriftzeichen und Alphabete aller Zeiten und aller Völker des Erdkreises*. Druck und Verlag der kaiserlich-königlichen Hof- und Staatsdruckerei, Wien, 1880.

C.E. Godakumbura. *Catalogue of Ceylonese Manuscripts*. The Royal Library, Copenhagen, 1980.

Y. Haralambous and J. Plaice. "First Applications of Ω: Greek, Arabic, Khmer, Poetica, ISO 10646/UNICODE, etc.". In *Proceedings of the 15th TeX Users Group Annual Meeting (Santa Barbara). TUGboat*, **15** (3), pp. 344-352, 1994.

ISO. *Information technology — Universal Multiple-octet Coded Character Set*. ISO/IEC 10646-1:1993(e) edition, 1993.

H.M. Lambert *Introduction to the Scripts of South India and Ceylon, manuscript prepared as a companion to: Introduction to the Devanagari Script, for Students of Sanskrit, Hindi, Marathi, Gujarati and Bengali*. Oxford University Press, 1983.

J. Levine, T. Mason, and D. Brown. *lex & yacc*. O'Reilly & Associates, Inc., Sebastopol, California, 1992.

A. Nakanishi. *Writing systems of the World*. Charles E. Tuttle Company, Tokyo, 1980.

J. Plaice. "Progress in the Ω Project". In *Proceedings of the 15th TeX Users Group Annual Meeting (Santa Barbara)*. 1994. *TUGboat*, **15** (3), pp. 320-324, 1994.

K.D. Somadasa. *Catalogue of the Sinhalese Manuscripts in the Wellcome Institute for the History of Medicine*. Wellcome Institute, London, 1994.

C. Thiele. "TeX, Linguistics and Journal Production". In *TeX Users Group Eighth Annual Meeting, Seattle, August 24–26, 1987*. 1987.

D. Wujastyk. "Standardization of Romanized Sanskrit for Electronic Data Transfer and Screen Representation". *Sesame Bulletin*, **4**(1), 27–29, 1991.

◇ Yannis Haralambous
187, rue Nationale
59800 Lille, France.
Email: `haralambous@univ-lille1.fr`

Table 3: Sinhalese consonants and vowel combinations

Part a. Without vowel, and vowels 'a'–'r̥̄'											
		a	ā	ä	ǟ	i	ī	u	ū	ṛ	ṝ
		ෘ	ෘා	ෘැ	ෘෑ	ෟ	ෟ්	ෳ	ෳෟ	ඍ	ඎ
ka	ක්	ක	කා	කැ	කෑ	කි	කී	කු	කූ	කෘ	කෲ
kha	ඛ්	ඛ	ඛා	ඛැ	ඛෑ	ඛි	ඛී	ඛු	ඛූ	ඛෘ	ඛෲ
ga	ග්	ග	ගා	ගැ	ගෑ	ගි	ගී	ගු	ගූ	ගෘ	ගෲ
gha	ඝ්	ඝ	ඝා	ඝැ	ඝෑ	ඝි	ඝී	ඝු	ඝූ	ඝෘ	ඝෲ
ṅa	ඞ්	ඞ	ඞා	ඞැ	ඞෑ	ඞි	ඞී	ඞු	ඞූ	ඞෘ	ඞෲ
ca	ච්	ච	චා	චැ	චෑ	චි	චී	චු	චූ	චෘ	චෲ
cha	ඡ්	ඡ	ඡා	ඡැ	ඡෑ	ඡි	ඡී	ඡු	ඡූ	ඡෘ	ඡෲ
ja	ජ්	ජ	ජා	ජැ	ජෑ	ජි	ජී	ජු	ජූ	ජෘ	ජෲ
jha	ඣ්	ඣ	ඣා	ඣැ	ඣෑ	ඣි	ඣී	ඣු	ඣූ	ඣෘ	ඣෲ
ña	ඤ්	ඤ	ඤා	ඤැ	ඤෑ	ඤි	ඤී	ඤු	ඤූ	ඤෘ	ඤෲ
ṭa	ට්	ට	ටා	ටැ	ටෑ	ටි	ටී	ටු	ටූ	ටෘ	ටෲ
ṭha	ඨ්	ඨ	ඨා	ඨැ	ඨෑ	ඨි	ඨී	ඨු	ඨූ	ඨෘ	ඨෲ
ḍa	ඩ්	ඩ	ඩා	ඩැ	ඩෑ	ඩි	ඩී	ඩු	ඩූ	ඩෘ	ඩෲ
ḍha	ඪ්	ඪ	ඪා	ඪැ	ඪෑ	ඪි	ඪී	ඪු	ඪූ	ඪෘ	ඪෲ
ṇa	ණ්	ණ	ණා	ණැ	ණෑ	ණි	ණී	ණු	ණූ	ණෘ	ණෲ
ta	ත්	ත	තා	තැ	තෑ	ති	තී	තු	තූ	තෘ	තෲ
tha	ථ්	ථ	ථා	ථැ	ථෑ	ථි	ථී	ථු	ථූ	ථෘ	ථෲ
da	ද්	ද	දා	දැ	දෑ	දි	දී	දු	දූ	දෘ	දෲ
dha	ධ්	ධ	ධා	ධැ	ධෑ	ධි	ධී	ධු	ධූ	ධෘ	ධෲ
na	න්	න	නා	නැ	නෑ	නි	නී	නු	නූ	නෘ	නෲ
pa	ප්	ප	පා	පැ	පෑ	පි	පී	පු	පූ	පෘ	පෲ
pha	ඵ්	ඵ	ඵා	ඵැ	ඵෑ	ඵි	ඵී	ඵු	ඵූ	ඵෘ	ඵෲ
ba	බ්	බ	බා	බැ	බෑ	බි	බී	බු	බූ	බෘ	බෲ
bha	භ්	භ	භා	භැ	භෑ	භි	භී	භු	භූ	භෘ	භෲ
ma	ම්	ම	මා	මැ	මෑ	මි	මී	මු	මූ	මෘ	මෲ
ya	ය්	ය	යා	යැ	යෑ	යි	යී	යු	යූ	යෘ	යෲ
ra	ර්	ර	රා	ඍ	ඎ	රි	රී	රු	රූ	රෘ	රෲ
la	ල්	ල	ලා	ලැ	ලෑ	ලි	ලී	ලු	ලූ	ලෘ	ලෲ
va	ව්	ව	වා	වැ	වෑ	වි	වී	වු	ෟ	වෘ	වෲ
śa	ශ්	ශ	ශා	ශැ	ශෑ	ශි	ශී	ශු	ශූ	ශෘ	ශෲ
ṣa	ෂ්	ෂ	ෂා	ෂැ	ෂෑ	ෂි	ෂී	ෂු	ෂූ	ෂෘ	ෂෲ
sa	ස්	ස	සා	සැ	සෑ	සි	සී	සු	සූ	සෘ	සෲ
ha	හ්	හ	හා	හැ	හෑ	හි	හී	හු	හූ	හෘ	හෲ
ḷa	ළ්	ළ	ළා	ළැ	ළෑ	ළි	ළී	ළු	ළූ	ළෘ	ළෲ
fa	ෆ්	ෆ	ෆා	ෆැ	ෆෑ	ෆි	ෆී	ෆු	ෆූ	ෆෘ	ෆෲ

Part b. Vowels ‘ḷ’–‘au’, anusvara, visarga										
	ḷ	ḹ	e	ē	ai	o	ō	au	aṃ	aḥ
	ඔ	ඕ	එ	ඒ	ඓ	ඔ	ඕ	ඖ	ං	ඃ
ka	කෘ	කෲ	කෙ	කේ	කෛ	කො	කෝ	කෞ	කං	කඃ
kha	ඛෘ	ඛෲ	ඛෙ	ඛේ	ඛෛ	ඛො	ඛෝ	ඛෞ	ඛං	ඛඃ
ga	ගෘ	ගෲ	ගෙ	ගේ	ගෛ	ගො	ගෝ	ගෞ	ගං	ගඃ
gha	ඝෘ	ඝෲ	ඝෙ	ඝේ	ඝෛ	ඝො	ඝෝ	ඝෞ	ඝං	ඝඃ
ṅa	ඞෘ	ඞෲ	ඞෙ	ඞේ	ඞෛ	ඞො	ඞෝ	ඞෞ	ඞං	ඞඃ
ca	චෘ	චෲ	චෙ	චේ	චෛ	චො	චෝ	චෞ	චං	චඃ
cha	ඡෘ	ඡෲ	ඡෙ	ඡේ	ඡෛ	ඡො	ඡෝ	ඡෞ	ඡං	ඡඃ
ja	ජෘ	ජෲ	ජෙ	ජේ	ජෛ	ජො	ජෝ	ජෞ	ජං	ජඃ
jha	ඣෘ	ඣෲ	ඣෙ	ඣේ	ඣෛ	ඣො	ඣෝ	ඣෞ	ඣං	ඣඃ
ña	ඤෘ	ඤෲ	ඤෙ	ඤේ	ඤෛ	ඤො	ඤෝ	ඤෞ	ඤං	ඤඃ
ṭa	ටෘ	ටෲ	ටෙ	ටේ	ටෛ	ටො	ටෝ	ටෞ	ටං	ටඃ
ṭha	ඨෘ	ඨෲ	ඨෙ	ඨේ	ඨෛ	ඨො	ඨෝ	ඨෞ	ඨං	ඨඃ
ḍa	ඩෘ	ඩෲ	ඩෙ	ඩේ	ඩෛ	ඩො	ඩෝ	ඩෞ	ඩං	ඩඃ
ḍha	ඪෘ	ඪෲ	ඪෙ	ඪේ	ඪෛ	ඪො	ඪෝ	ඪෞ	ඪං	ඪඃ
ṇa	ණෘ	ණෲ	ණෙ	ණේ	ණෛ	ණො	ණෝ	ණෞ	ණං	ණඃ
ta	තෘ	තෲ	තෙ	තේ	තෛ	තො	තෝ	තෞ	තං	තඃ
tha	ථෘ	ථෲ	ථෙ	ථේ	ථෛ	ථො	ථෝ	ථෞ	ථං	ථඃ
da	දෘ	දෲ	දෙ	දේ	දෛ	දො	දෝ	දෞ	දං	දඃ
dha	ධෘ	ධෲ	ධෙ	ධේ	ධෛ	ධො	ධෝ	ධෞ	ධං	ධඃ
na	නෘ	නෲ	නෙ	නේ	නෛ	නො	නෝ	නෞ	නං	නඃ
pa	පෘ	පෲ	පෙ	පේ	පෛ	පො	පෝ	පෞ	පං	පඃ
pha	ඵෘ	ඵෲ	ඵෙ	ඵේ	ඵෛ	ඵො	ඵෝ	ඵෞ	ඵං	ඵඃ
ba	බෘ	බෲ	බෙ	බේ	බෛ	බො	බෝ	බෞ	බං	බඃ
bha	භෘ	භෲ	භෙ	භේ	භෛ	භො	භෝ	භෞ	භං	භඃ
ma	මෘ	මෲ	මෙ	මේ	මෛ	මො	මෝ	මෞ	මං	මඃ
ya	යෘ	යෲ	යෙ	යේ	යෛ	යො	යෝ	යෞ	යං	යඃ
ra	රෘ	රෲ	රෙ	රේ	රෛ	රො	රෝ	රෞ	රං	රඃ
la	ලෘ	ලෲ	ලෙ	ලේ	ලෛ	ලො	ලෝ	ලෞ	ලං	ලඃ
va	වෘ	වෲ	වෙ	වේ	වෛ	වො	වෝ	වෞ	වං	වඃ
śa	ශෘ	ශෲ	ශෙ	ශේ	ශෛ	ශො	ශෝ	ශෞ	ශං	ශඃ
ṣa	ෂෘ	ෂෲ	ෂෙ	ෂේ	ෂෛ	ෂො	ෂෝ	ෂෞ	ෂං	ෂඃ
sa	සෘ	සෲ	සෙ	සේ	සෛ	සො	සෝ	සෞ	සං	සඃ
ha	හෘ	හෲ	හෙ	හේ	හෛ	හො	හෝ	හෞ	හං	හඃ
ḷa	ළෘ	ළෲ	ළෙ	ළේ	ළෛ	ළො	ළෝ	ළෞ	ළං	ළඃ
fa	ෆෘ	ෆෲ	ෆෙ	ෆේ	ෆෛ	ෆො	ෆෝ	ෆෞ	ෆං	ෆඃ

Table 4: Table of Devanagari, Tamil, Malayalam and Sinhalese characters and the different input modes

NAME	D	T	M	S	CSX	SEVENBIT	LATEX
ANUSVARA	ं	◌	ം	ං	m, M	.m, M	\d{m}, \d{M}
VISARGA	ः	◌ः	ഃ	ඃ	h, H	.h, H	\d{h}, \d{H}
VOW. A	अ	அ	അ	අ	a, A	a	a, A
VOW. AA	आ	ஆ	ആ	ආ	ā, Ā	aa, A	\={a}, \={A}
VOW. AAA	-	-	-	ඇ	ä, Ä	*"a*, *.A*	\"{a}, \"{A}
VOW. AAAA	-	-	-	ඈ	ā̈, Ā̈	*"aa*, *.AA*	\diatop[\=|\"a], \diatop[\=|\"A]
VOW. I	इ	இ	ഇ	ඉ	i, I	i	i, I
VOW. II	ई	ஈ	ഈ	ඊ	ī, Ī	ii I	\={\i}, \={i} \={I}
VOW. U	उ	உ	ഉ	උ	u, U	u	u, U
VOW. UU	ऊ	ஊ	ഊ	ඌ	ū, Ū	uu, U	\={u}, \={U}
VOW. VOC. R	ऋ	-	ഋ	ඍ	ṛ, Ṛ	.r	\d{r}, \d{R}
VOW. VOC. RR	ॠ	-	ൠ	ඎ	ṝ, Ṝ	.R	\diatop[\=|\d r], \diatop[\=|\d r]
VOW. VOC. L	ऌ	-	ഌ	ඏ	ḷ, Ḷ	.l	\d{l}, \d{L}
VOW. VOC. LL	ॡ	-	ൡ	ඐ	ḹ, Ḹ	.L	\diatop[\=|\d l], \diatop[\=|\d L]
VOW. CANDRA E	ऍ	-	-	-	ĕ	??!!	\u{e}
VOW. SHORT E	ऎ	எ	എ	එ	ĕ, Ĕ	^e	\v{e}, \v{E}
VOW. E	ए	ஏ	ഏ	ඒ	e, E	e	e, E
VOW. AI	ऐ	ஐ	ഐ	ඓ	ai, Ai, AI	ai, E	ai, Ai, AI
VOW. CANDRA O	ऑ	-	-	-	ŏ	??!!	\u{o}
VOW. SHORT O	ऒ	ஒ	ഒ	ඔ	ŏ, Ŏ	^o	\v{o}, \v{O}
VOW. O	ओ	ஓ	ഓ	ඕ	o, O	o	o, O
VOW. AU	औ	ஔ	ഔ	ඖ	au, Au, AU	au, O	au, Au, AU
CONS. KA	क	க	ക	ක	k, K	k	k, K
CONS. KHA	ख	-	ഖ	ඛ	kh, Kh, KH	kh, K	kh, Kh, KH
CONS. GA	ग	-	ഗ	ග	g, G	g	g, G
CONS. GHA	घ	-	ഘ	ඝ	gh, Gh, GH	gh, G	gh, Gh, GH
CONS. NGA	ङ	ங	ങ	ඟ	ṅ, Ṅ	"n	\.{n}, \.{N}
CONS. CA	च	ச	ച	ච	c, C	c	c, C
CONS. CHA	छ	-	ഛ	ඡ	ch, Ch, CH	ch, C	ch, Ch, CH
CONS. JA	ज	ஜ	ജ	ජ	j, J	j	j, J
CONS. JHA	झ	-	ഝ	ඣ	jh, Jh, JH	jh, J	jh, Jh, JH
CONS. NYA	ञ	ஞ	ഞ	ඤ	ñ, Ñ	~n	\~{n}, \~{N}
CONS. TTA	ट	ட	ട	ට	ṭ, Ṭ	.t	\d{t}, \d{T}
CONS. TTHA	ठ	-	ഠ	ඨ	ṭh, Ṭh, ṬH	.th, .T	\d{t}h, \d{T}h, \d{T}H
CONS. DDA	ड	-	ഡ	ඩ	ḍ, Ḍ	.d	\d{d}, \d{D}
CONS. DDHA	ढ	-	ഢ	ඪ	ḍh, Ḍh, ḌH	.dh, .D	\d{d}h, \d{D}h, \d{D}H
CONS. NNA	ण	ண	ണ	ණ	ṇ, Ṇ	.n	\d{n}, \d{N}

NAME	D	T	M	S	CSX	SEVENBIT	LATEX
CONS. TA	त	த	ത	ත	t, T	t	t, T
CONS. THA	थ	-	ഥ	ථ	th, Th, TH	th, T	th, Th, TH
CONS. DA	द	-	ദ	ද	d, D	d	d, D
CONS. DHA	ध	-	ധ	ධ	dh, Dh, DH	dh, D	dh, Dh, DH
CONS. NA	न	ந	ന	න	n, N	n	n, N
CONS. NNNA	ऩ	ன	-	-	**??!!**	??!!	??!!
CONS. PA	प	ப	പ	ප	p, P	p	p, P
CONS. PHA	फ	-	ഫ	ඵ	ph, Ph, PH	ph, P	ph, Ph, PH
CONS. BA	ब	-	ബ	බ	b, B	b	b, B
CONS. BHA	भ	-	ഭ	භ	bh, Bh, BH	bh, B	bh, Bh, BH
CONS. MA	म	ம	മ	ම	m, M	m	m, M
CONS. YA	य	ய	യ	ය	y, Y	y	y, Y
CONS. RA	र	ர	ര	ර	r, R	r	r, R
CONS. RRA	ऱ	ற	റ	-	**??!!**	*"r*	??!!
CONS. LA	ल	ல	ല	ල	l, L	l	l, L
CONS. LLA	ळ	ள	ള	ළ	l̲	L	\b{l}, \b{L}
CONS. LLLA	ऴ	ழ	ഴ	-	**??!!**	*"l*	??!!
CONS. VA	व	வ	വ	ව	v, V	v	v, V
CONS. SHA	श	ஶ	ശ	ශ	ś, Ś	"s	\'s, \'S
CONS. SSA	ष	-	ഷ	ෂ	ṣ, Ṣ	.s	\d{s}, \d{S}
CONS. SA	स	ஸ	സ	ස	s, S	s	s, S
CONS. HA	ह	ஹ	ഹ	හ	h, H	h	h, H
CONS. FA	फ़	-	-	ෆ	f, F	f	f, F
CONS. NAS. GA	-	-	-	ඟ	n̆g, N̆g, N̆G	*Ng*	\u{n}g, \u{N}g, \u{N}G
CONS. NAS. CA	-	-	-	ඤ	n̆c, N̆c, N̆C	*Nc*	\u{n}c, \u{N}c, \u{N}C
CONS. NAS. DDA	-	-	-	ඬ	n̆ḍ, N̆ḍ, N̆Ḍ	*N.d*	\u{n}\d{d}, \u{N}\d{d}, \u{N}\d{D}
CONS. NAS. DA	-	-	-	ඳ	n̆d, N̆d, N̆D	*Nd*	\u{n}d, \u{N}d, \u{N}D
CONS. NAS. BA	-	-	-	ඹ	m̆b, M̆b, M̆B	*Nb*	\u{m}b, \u{M}b, \u{M}B
CONS. NAS. JA	-	-	-	ඦ	n̆j, N̆j, N̆J	*Nj*	\u{n}j, \u{N}j, \u{N}J
CONS. QA	क़	-	-	-	q	q	q
CONS. KHHA	ख़	-	-	-	k̲h	.kh, .K	\b{k}\b{h}
CONS. GHHA	ग़	-	-	-	g̲	.g	\b{g}
CONS. ZA	ज़	-	-	-	z	z	z
CONS. DDDHA	ड़	-	-	-	r̲	R	\b{r}
CONS. RHA	ढ़	-	-	-	r̲h	Rh	\b{r}h
CONS. YYA	य़	-	-	-	**??!!**	*"y*	??!!

NOTES:

— Columns D, T, M, S, stand respectively for Devanagari, Tamil, Malayalam and Sinhalese. The fonts used in this paper for the first three scripts have been made by Frans Velthuis (velthuis@rc.rug.nl), Thomas Ridgeway (165 McGraw Street, Seattle, WA 98109 USA), Jeroen Hellingman (jhelling@cs.ruu.nl) and the author. Some of them are still under β-status, so please contact their respective authors for more information on their availability.

— SEVENBIT column: Entries in slanted style are extensions to Frans Velthuis' transcription, proposed by the author.