The akshar package

Vu Van Dung

Version 0.1 - 2020/05/17

Abstract

This package provides tools to deal with special characters in a Devanagari string.

Contents

1	Introduction	1
2	User manual 2.1 \LaTeX 2.2 expl3 functions	1 1 2
3		3 4 5
Inde		8

1 Introduction

When dealing with processing strings in the Devanagari script, normal \LaTeX commands usually find some difficulties in distinguishing "normal" characters, like क, and "special" characters, for example ् or ी. Let's consider this example code:

- 1 \ExplSyntaxOn
- 2 \tl set:Nn \l tmpa tl { की}

2 tokens.

- 3 \tl count:N \l tmpa tl \c space token tokens.
- 4 \ExplSyntaxOff

The output is 2, but the number of characters in it is only one! The reason is quite simple: the compiler treats as a normal character, and it shouldn't do so.

To tackle that, this package provides expl3 functions to "convert" a given string, written in the Devanagari script, to a sequence of token lists. each of these token lists is a "true" Devanagari character. You can now do anything you want with this sequence; and this package does provide some front-end macros for some simple actions on the input string.

2 User manual

2.1 $\text{LT}_{F}X 2_{\varepsilon}$ macros

\aksharStrLen

 $\arrowvert aksharStrLen {\langle token list \rangle}$

Return the number of Devanagari characters in the (token list).

There are 4 characters in नमस्कार. expl3 returns 7, which is wrong.

- ा There are \aksharStrLen{ नमस्कार} characters in नमस्कार.\par
- 2 \ExplSyntaxOn
- $3 \neq \text{pkg}\{\text{expl3}\}\$ returns $\sim \text{tl}\$ count:n $\{ \ \$ नमस्कार $\}, \sim \text{which}\$ is $\sim \text{wrong}.$
- 4 \ExplSyntaxOff

\aksharStrChar

 $\arrowvert aksharStrChar {\langle token list \rangle} {\langle n \rangle}$

Return the n-th character of the token list.

3rd character of नमस्कार is स्का. It is not स.

- ा 3rd character of नमस्कारांs \aksharStrChar{ नमस्कार}{3}.\par
- 2 \ExplSyntaxOn
- 3 It~is~not~\tl item:nn { नमस्कार} {3}.
- 4 \ExplSyntaxOff

\aksharStrReplace \aksharStrReplace* \aksharStrReplace $\{\langle tl\ 1\rangle\}\ \{\langle tl\ 2\rangle\}\ \{\langle tl\ 3\rangle\}$

Replace all occurences of $\langle tl \ 2 \rangle$ in $\langle tl \ 1 \rangle$ with $\langle tl \ 3 \rangle$, and leaves the modified $\langle tl \ 1 \rangle$ in the input stream.

The starred variant will replace only the first occurence of $\langle tl \ 2 \rangle$, all others are left intact.

expl3 output: स्कास्कास्काडडस्कांळीस्कास्काड \aksharStrReplace output: स्कास्कास्काडडमंळीस्कास्काड

- ∖ExplSyntaxOn
- $_2 \pkg{expl3} \sim output:\par$
- ₃ \tl_set:Nn \l_tmpa_tl { मममडडमंळीममड}
- 4 \tl replace all:Nnn \l tmpa tl { म} { स्का}
- 5 \tl use:N \l tmpa tl\par
- $_{6} \cs{aksharStrReplace} \sim output:\par$
- ७ \aksharStrReplace { मममडडमंळीममड} { म} { स्का}
- « \ExplSyntaxOff

2.2 expl3 functions

This section assumes that you have a basic knowledge in LATEX3 programming. All macros in 2.1 directly depend on the following function, so it is much more powerful than all features we have described above.

\akshar_convert:Nn \akshar_convert:(cn|Nx|cx) \akshar_convert:Nn \langle seq var \rangle \langle \taken list \rangle \rangle

This function converts $\langle token \ list \rangle$ to a sequence of characters, that sequence is stored in $\langle seq \ var \rangle$. The assignment to $\langle seq \ var \rangle$ is local to the current TeX group.

न, म, स्का, and र

- $_1$ \ExplSyntaxOn
- 2 \akshar convert:Nn \l tmpa seq { नमस्कार}
- 3 \seq use:Nnnn \l tmpa seq { ~and~ } { ,~ } { ,~and~ }
- 4 \ExplSyntaxOff

3 Implementation

- ₁ ⟨@@=akshar⟩
- 2 (*package)

Declare the package. By loading fontspec, xparse, and in turn, expl3, are also loaded.

```
3 \RequirePackage{fontspec}
4 \ProvidesExplPackage {akshar} {2020/05/17} {0.1}
5 {Support for syllables in the Devanagari script (JV)}
```

3.1 Variable declarations

\c__akshar_joining_tl
\c__akshar_diacritics_tl

 $\verb|\c_akshar_joining_t| These variables store the special characters we need to take into account:$

- \c_akshar_joining_tl is the "connecting" character o.

(End definition for \c_akshar_joining_tl and \c_akshar_diacritics_tl.)

\l__akshar_prev_joining_bool

When we get to a normal character, we need to know whether it is joined, i.e. whether the previous character is the joining character. This boolean variable takes care of that.

```
\bool_new:N \l__akshar_prev_joining_bool
(End definition for \l__akshar_prev_joining_bool.)
```

\l_akshar_char_seq This local sequence stores the output of the converter.

```
14 \seq_new:N \l__akshar_char_seq
(End definition for \l__akshar_char_seq.)
```

```
\l_akshar_tmpa_tl
\l_akshar_tmpb_tl
\l_akshar_tmpb_seq
\l_akshar_tmpb_seq
\l_akshar_tmpd_seq
\l_akshar_tmpd_seq
\l_akshar_tmpe_seq
\l_akshar_tmpe_seq
\l_akshar_tmpa_int
```

Some temporary variables.

```
15 \tl_new:N \l_akshar_tmpa_tl
16 \tl_new:N \l_akshar_tmpb_tl
17 \seq_new:N \l_akshar_tmpa_seq
18 \seq_new:N \l_akshar_tmpb_seq
19 \seq_new:N \l_akshar_tmpc_seq
20 \seq_new:N \l_akshar_tmpd_seq
21 \seq_new:N \l_akshar_tmpe_seq
22 \int_new:N \l_akshar_tmpa_int
```

(End definition for $\l_{-akshar_tmpa_tl}$ and others.)

3.2 Messages

In \akshar_convert: Nn and friends, the argument needs to be a sequence variable. There will be an error if it isn't.

In \aksharStrChar, we need to guard against accessing an 'out-of-bound' character (like trying to get the 8th character in a 5-character string.)

3.3 Utilities

\tl_if_in:No<u>TF</u> When we get to a character which is not the joining one, we need to know if it is a diacritic. The current character is stored in a variable, so an expanded variant is needed. We only need it to expand only once.

```
42 \prg_generate_conditional_variant:Nnn \tl_if_in:Nn { No } { TF }
(End definition for \tl_if_in:NoTF.)
```

\seq_set_split:Nxx A variant we will need in __akshar_var_if_global.

43 \cs_generate_variant:Nn \seq_set_split:Nnn { Nxx }

(End definition for \seq_set_split:Nxx.)

 $\mbox{msg_error:nnx}$ Some variants of l3msg functions that we will need when issuing error mes- $\mbox{msg_error:nnnxx}$ sages.

```
44 \cs_generate_variant:Nn \msg_error:nnn { nnx }
45 \cs_generate_variant:Nn \msg_error:nnnnn { nnnxx }
```

(End definition for $\mbox{msg_error:nnx}$ and $\mbox{msg_error:nnnxx.}$)

__akshar_var_if_global:N<u>TF</u>
\c__akshar_str_g_tl
\c__akshar_str_seq_tl

This conditional checks if #1 is a global sequence variable or not. In other words, it returns true iff #1 is a control sequence in the format g_{α} ame geq. If it is not a sequence variable, this function will (TODO) issue an error message.

```
46 \tl_const:Nx \c__akshar_str_g_tl { \tl_to_str:n {g} }
47 \tl_const:Nx \c__akshar_str_seq_tl { \tl_to_str:n {seq} }
48 \prg_new_conditional:Npnn \__akshar_var_if_global:N #1 { T, F, TF }
49
    {
      \bool_if:nTF
50
        { \exp_last_unbraced:Nf \use_iii:nnn { \cs_split_function:N #1 } }
51
        {
          \msg_error:nnx { akshar } { err_not_a_sequence_variable }
            { \token_to_str:N #1 }
          \prg_return_false:
55
        }
56
        {
57
          \seq_set_split:Nxx \l__akshar_tmpb_seq { \token_to_str:N _ }
58
            { \exp_last_unbraced:Nf \use_i:nnn { \cs_split_function:N #1 } }
59
          \seq_get_left:NN \l__akshar_tmpb_seq \l__akshar_tmpa_tl
60
          \seq get right:NN \l akshar tmpb seg \l akshar tmpb tl
61
          \tl_if_eq:NNTF \c__akshar_str_seq_tl \l__akshar_tmpb_tl
62
            {
              \tl_if_eq:NNTF \c__akshar_str_g_tl \l__akshar_tmpa_tl
                { \prg_return_true: } { \prg_return_false: }
            }
            {
67
              \msg_error:nnx { akshar } { err_not_a_sequence_variable }
68
                { \token_to_str:N #1 }
69
              \prg_return_false:
70
72
        }
    }
```

(End definition for $_$ akshar_var_if_global:NTF, $_$ akshar_str_g_tl, and $_$ akshar_str_seq_tl.)

__akshar_int_append_ordinal:n Append st, nd, rd or th to interger #1. Will be needed in error messages.

```
74 \cs_new:Npn \__akshar_int_append_ordinal:n #1
 75
                     {
 76
                                      #1
                                      \int_case:nnF { #1 }
 77
 78
                                                 {
                                                              { 11 } { th }
 79
                                                              { 12 } { th }
 80
                                                             { 13 } { th }
 81
                                                              { -11 } { th }
 82
                                                              { -12 } { th }
 83
                                                              { -13 } { th }
 84
 85
                                                 }
 86
                                                  {
                                                              \int \int d^2 r 
 87
                                                                        {
 88
                                                                                     \int_case:nnF { #1 - 10 * (#1 / 10) }
 89
                                                                                                {
 90
                                                                                                           { 1 } { st }
 91
                                                                                                           { 2 } { nd }
 92
                                                                                                           { 3 } { rd }
 93
                                                                                                 } { th }
 94
                                                                        }
                                                                        {
                                                                                     \int_case:nnF { (- #1) - 10 * ((- #1) / 10) }
 97
 98
                                                                                                           { 1 } { st }
 99
                                                                                                           { 2 } { nd }
100
                                                                                                           { 3 } { rd }
101
102
                                                                                                 } { th }
                                                                        }
103
                                                 }
104
                          }
```

(End definition for $_$ akshar_int_append_ordinal:n.)

The \akshar_convert:Nn function and its variants

\akshar_convert:Nn \akshar_convert:cn \akshar_convert:Nx \akshar_convert:cx

This converts #2 to a sequence of true Devanagari characters. The sequence is set to #1, which should be a sequence variable. The assignment is local.

```
106 \cs_new:Npn \akshar_convert:Nn #1 #2
```

Clear anything stored in advance. We don't want different calls of the function to conflict with each other.

```
\seq_clear:N \l__akshar_char_seq
      \bool_set_false:N \l__akshar_prev_joining_bool
109
```

Loop through every token of the input.

```
\tl_map_variable:NNn {#2} \l__akshar_map_tl
111
        {
           \tl_if_in:NoTF \c__akshar_diacritics_tl {\l__akshar_map_tl}
```

It is a diacritic. We append the current diacritic to the last item of the sequence instead of pushing the diacritic to a new sequence item.

```
114
           \seq_put_right:Nx \l__akshar_char_seq
             { \l__akshar_tmpa_tl \l__akshar_map_tl }
116
         }
         {
118
           \tl_if_eq:NNTF \l__akshar_map_tl \c__akshar_joining_tl
119
             {
120
```

In this case, the character is the joining character, Q. What we do is similar to the above case, but \l_akshar_prev_joining_bool is set to true so that the next character is also appended to this item.

Now the character is normal. We see if we can push to a new item or not. It depends on the boolean variable.

```
\bool_if:NTF \l__akshar_prev_joining_bool
128
                       \seq pop right:NN \l akshar char seq \l akshar tmpa tl
129
                       \seq_put_right:Nx \l__akshar_char_seq
130
                         { \l_akshar_tmpa_tl \l_akshar_map_tl }
                       \bool_set_false:N \l__akshar_prev_joining_bool
                     }
                       \seq_put_right:Nx \l__akshar_char_seq { \l__akshar_map_tl }
136
                 }
             }
138
         }
139
```

Set #1 to \l_akshar_char_seq. The package automatically determines whether the variable is a global one or a local one.

Generate variants that might be helpful for some.

```
\text{\cs_generate_variant:Nn \akshar_convert:Nn { cn, Nx, cx }
```

(End definition for \akshar_convert:Nn. This function is documented on page 2.)

3.5 Other internal functions

__akshar_seq_push_seq:NN Append sequence #1 to the end of sequence #2. A simple loop will do.

```
145 \cs_new:Npn \__akshar_seq_push_seq:NN #1 #2
146 { \seq_map_inline:Nn #2 { \seq_put_right:Nn #1 { ##1 } } }

(End definition for \__akshar_seq_push_seq:NN.)
```

__akshar_replace_all:Nnnn

This function replaces all occurences of #3 in #2 by #4. The result is stored in a sequence variable #1.

The algorithm used in this function: We will use $\l_akshar_tmpa_int$ to store the "current position" in the sequence of #3. At first it is set to 1.

We will store any subsequence of #2 that may match #3 to a temporary sequence. If it doesn't match, we push this temporary sequence to the output, but if it matches, #4 is pushed instead.

We loop over #2. For each of these loops, we need to make sure the \l_- -akshar_tmpa_int-th item must indeed appear in #3. So we need to compare that with the length of #3.

If now \l_akshar_tmpa_int is greater than the length of #3, the whole
 of #3 has been matched somewhere, so we reinitialize the integer to 1
 and push #4 to the output.

Note that it is possible that the current character might be the start of another match, so we have to compare it to the first character of #3. If

they are not the same, we may now push the current mapping character to the output and proceed; otherwise the current character is pushed to the temporary variable.

- Otherwise, we compare the current loop character of #2 with the \l_-akshar_tmpa_int-th character of #3.
 - If they are the same, we still have a chance that it will match, so we increase the "iterator" \l_akshar_tmpa_int by 1 and push the current mapping character to the temporary sequence.
 - If they are the same, the temporary sequence won't match. Let's push that sequence to the output and set the iterator back to 1.
 Note that now the iterator has changed. Who knows whether the current character may start a match? Let's compare it to the first character of #3, and do as in the case of \l_akshar_tmpa_int is greater than the length of #3.

The complexity of this algorithm is $O(m \max(n, p))$, where m, n, p are the lengths of the sequences created from #2, #3 and #4. As #3 and #4 are generally short strings, this is (almost) linear to the length of the original sequence #2.

```
\cs_new:Npn \__akshar_replace_all:Nnnn #1 #2 #3 #4
149
       \akshar_convert:Nn \l__akshar_tmpc_seq {#2}
       \akshar_convert:Nn \l__akshar_tmpd_seq {#3}
       \akshar_convert:Nn \l__akshar_tmpe_seq {#4}
       \seq_clear:N \l__akshar_tmpa_seq
       \seq_clear:N \l__akshar_tmpb_seq
       \int_set:Nn \l__akshar_tmpa_int { 1 }
154
       \seq_map_variable:NNn \l__akshar_tmpc_seq \l__akshar_map_tl
155
156
           \int_compare:nNnTF
             { \l_akshar_tmpa_int } = { 1 + \seq_count:N \l_akshar_tmpd_seq }
158
             {
               \seq_clear:N \l__akshar_tmpb_seq
160
               \__akshar_seq_push_seq:NN \l__akshar_tmpa_seq \l__akshar_tmpe_seq
161
               \int_set:Nn \l__akshar_tmpa_int { 1 }
162
163
               \tl_set:Nx \l__akshar_tmpa_tl
                 { \seq_item:Nn \l__akshar_tmpd_seq { 1 } }
164
               \tl_if_eq:NNTF \l__akshar_map_tl \l__akshar_tmpa_tl
165
166
                 {
                   \int_incr:N \l__akshar_tmpa_int
167
                    \seq_put_right:NV \l__akshar_tmpb_seq \l__akshar_map_tl
168
                 { \seq_put_right:NV \l__akshar_tmpa_seq \l__akshar_map_tl }
             }
             {
               \tl_set:Nx \l__akshar_tmpa_tl
                 { \seq_item:Nn \l__akshar_tmpd_seq { \l__akshar_tmpa_int } }
174
               \tl_if_eq:NNTF \l__akshar_map_tl \l__akshar_tmpa_tl
176
                   \int_incr:N \l__akshar_tmpa_int
                   \seq_put_right:NV \l__akshar_tmpb_seq \l__akshar_map_tl
178
                 }
                   \int_set:Nn \l__akshar_tmpa_int { 1 }
181
                   \__akshar_seq_push_seq:NN \l__akshar_tmpa_seq \l__akshar_tmpb_seq
                   \seq_clear:N \l__akshar_tmpb_seq
183
                   \tl_set:Nx \l__akshar_tmpa_tl
184
                     { \seq_item:Nn \l__akshar_tmpd_seq { 1 } }
185
                   \tl_if_eq:NNTF \l__akshar_map_tl \l__akshar_tmpa_tl
186
                     {
187
                        \int_incr:N \l__akshar_tmpa_int
188
                        \seq_put_right:NV \l__akshar_tmpb_seq \l__akshar_map_tl
                     { \seq_put_right:NV \l__akshar_tmpa_seq \l__akshar_map_tl }
                 }
             }
193
         }
194
```

```
\__akshar_seq_push_seq:NN \l__akshar_tmpa_seq \l__akshar_tmpb_seq
               196
                      \__akshar_var_if_global:NTF #1
                        { \seq_gset_eq:NN #1 \l__akshar_tmpa_seq }
               197
                        { \seq_set_eq:NN #1 \l__akshar_tmpa_seq }
               198
               199
               (End definition for \__akshar_replace_all:Nnnn.)
               3.6 Front-end \Delta T_E X 2_{\varepsilon} macros
\aksharStrLen Expands to the length of the string.
               200 \NewExpandableDocumentCommand \aksharStrLen {m}
               201
                      \akshar_convert:Nn \l__akshar_tmpa_seq {#1}
               202
                      \seq_count:N \l__akshar_tmpa_seq
               203
               204
               (End definition for \aksharStrLen. This function is documented on page 1.)
              Returns the n-th character of the string.
                  \NewExpandableDocumentCommand \aksharStrChar {mm}
               206
                      \akshar_convert:Nn \l__akshar_tmpa_seq {#1}
               207
                      \bool_if:nTF
               208
                        {
               209
                          \int \int d^2 x dx dx = 0 
                          211
                        { \seq_item:Nn \l__akshar_tmpa_seq { #2 } }
               214
                        {
                          \msg_error:nnnxx { akshar } { err_character_out_of_bound }
                            { #1 } { \__akshar_int_append_ordinal:n { #2 } }
               216
                            { \int_eval:n { 1 + \seq_count:N \l__akshar_tmpa_seq } }
               217
                          \scan_stop:
               218
                        }
               219
                    }
               (End definition for \aksharStrChar. This function is documented on page 2.)
              Replace any apparence of #3 of a string #2 with another string #4.
                  \NewExpandableDocumentCommand \aksharStrReplace {smmm}
```

\aksharStrReplace

\aksharStrChar

```
\IfBooleanTF {#1}
     { \iow_log:n { Do ~ nothing } }
224
     \seq_use:Nn \l__akshar_tmpa_seq {}
226
```

(End definition for \aksharStrReplace. This function is documented on page 2.)

228 (/package)

Index

The italic numbers denote the pages where the corresponding entry is described, numbers underlined point to the definition, all others indicate the places where it is used.

```
akshar internal commands:
akshar commands:
                                            \l__akshar_char_seq .....
                                                ...... 6, <u>14</u>, 108, 114, 115,
   \akshar_convert:Nn ..... 1, 2,
      3, 5, 106, 149, 150, 151, 202, 207
                                                121, 122, 129, 130, 135, 141, 142
```

$c_akshar_diacritics_tl$. 3, <u>6</u> , 112	\int_incr:N 167, 177, 188
<pre>\akshar_int_append_ordinal:n</pre>	\int_new:N 22
	\int_set:Nn 154, 162, 181
\cakshar_joining_tl 3, <u>6</u> , 119	iow commands:
\lakshar_map_tl 110, 112, 116,	\iow_log:n 224
119, 123, 131, 135, 155, 165,	
168, 170, 175, 178, 186, 189, 191	M
\l_akshar_prev_joining_bool	msg commands:
5, 13, 109, 124, 127, 132	\msg_error:nnn <u>44</u> , 44, 53, 68
_akshar_replace_all:Nnnn 147, 225	\msg_error:nnnnn 44, 45, 215
_akshar_seq_push_seq:NN	\msg_new:nnnn 23, 33
	(9
	N
\cakshar_str_g_tl <u>46</u>	\NewExpandableDocumentCommand
\cakshar_str_seq_tl <u>46</u>	200, 205, 221
\l_akshar_tmpa_int 6, 7, <u>15</u> , 154,	,,,,,
158, 162, 167, 174, 177, 181, 188	P
\l_akshar_tmpa_seq	prg commands:
<u>15</u> , 152, 161, 170,	\prg_generate_conditional
182, 191, 195, 197, 198, 202,	variant:Nnn 42
203, 207, 211, 213, 217, 225, 226	\prg_new_conditional:Npnn 48
\lakshar_tmpa_tl 15 ,	
60, 64, 114, 116, 121, 123, 129,	\prg_return_false: 55, 65, 70
131, 163, 165, 173, 175, 184, 186	\prg_return_true: 65
\l_akshar_tmpb_seq	\ProvidesExplPackage 4
160, 168, 1 78 , 182, 183, 189, 195	R
\lakshar_tmpb_tl 15, 61, 62	\RequirePackage 3
\lakshar_tmpc_seq <u>15</u> , 149, 155	0
\lakshar_tmpd_seq	S
<u>15</u> , 150, 158, 164, 174, 185	scan commands:
\l_akshar_tmpe_seq <u>15</u> , 151, 161	\scan_stop: 40, 218
\akshar_var_if_global 4	seq commands:
	\seq_clear:N 108, 152, 153, 160, 183
\akshar_var_if_global:NTF	\seq_count:N 158, 203, 211, 217
	\seq_get_left:NN60
\aksharStrChar 2, 3, <u>205</u>	\seq_get_right:NN 61
\aksharStrLen 1, <u>200</u>	\seq_gset_eq:NN 141, 197
\aksharStrReplace 2, <u>221</u>	\seq_item:Nn 164, 174, 185, 213
\aksharStrReplace* 2	\seq_map_inline:Nn 146
	\seq_map_variable:NNn 155
В	\seq_new:N 14, 17, 18, 19, 20, 21
bool commands:	\seq_pop_right:NN 114, 121, 129
\bool_if:NTF 127	\seq_put_right:Nn 115, 122, 130,
\bool_if:nTF 50, 208	135, 146, 168, 170, 178, 189, 191
\bool_new:N 13	\seq_set_eq:NN 142, 198
\bool_set_false:N 109, 132	\seq_set_split:Nnn 43, 43, 58
\bool_set_true:N 124	\seq_use:\Nn \frac{43}{256}, \frac{43}{36}
(5001_501_1140111	\seq_use:\nn 220
С	T
cs commands:	tl commands:
\cs_generate_variant:Nn	
	\tl_const:Nn 6, 7, 46, 47 \tl_if_eq:NNTF
\cs_new:Npn 74, 106, 145, 147	
	62, 64, 119, 165, 175, 186
\cs_split_function:N 51, 59	\tl_if_in:Nn 42
F	\tl_if_in:NnTF <u>42</u> , 112
E	\tl_map_variable:NNn 110
exp commands:	\tl_new:N 15, 16
\exp_last_unbraced:Nf 51, 59	\tl_set:Nn 163, 173, 184
	\tl_to_str:n 46, 47
I	token commands:
\IfBooleanTF 223	\token_to_str:N 40, 54, 58, 69
	\token_to_sti .n 40, 54, 56, 09
int commands:	(token_to_str.w 40, 54, 50, 03
\int_case:nnTF 77, 89, 97	U
\int_case:nnTF 77, 89, 97	U use commands: \use_i:nnn
\int_case:nnTF 77, 89, 97 \int_compare:nNnTF 87, 157	U use commands: