TIPITAKA-XML ANALYZING REPORT
(J.R. Bhaddacak, 2024-12-18)

0. Character-stat of Devanagari sources.

Character	Unicode	Frequency
	0020	9655934
◌्	094D	7346011
त	0924	6136199
◌ि	093F	5411693
◌ा	093E	5127234
स	0938	4251388
न	0928	3913155
प	092A	2788378
व	0935	2569503
◌े	0947	2486753
म	092E	2473040
◌ं	0902	2195417
क	0915	2098458
◌ो	094B	2042967
◌ु	0941	1786280
य	092F	1764703
र	0930	1623126
द	0926	1589685
च	091A	1480526
ह	0939	1147193
	200D	1083488
	000D	1051145
अ	0905	1010105
।	0964	958275
,	002C	943018
ग	0917	909660
ब	092C	829765
थ	0925	729173
भ	092D	727783
ज	091C	724880
ञ	091E	706012
◌ी	0940	683316
ध	0927	676754
ण	0923	652995
ख	0916	598054
‘	2018	590589
ल	0932	587255
ट	091F	540438
’	2019	518918
ठ	0920	364756
आ	0906	323092
◌ू	0942	292739
ए	090F	279860
उ	0909	273821
छ	091B	200935
इ	0907	200536

ङ	0919	182965
.	002E	172137
॰	0970	143306
…	2026	124776
1	0031	105447
–	2013	87922
झ	091D	82617
घ	0918	81998
॥	0965	80294
(	0028	78611
)	0029	78595
ड	0921	74872
2	0032	74156
फ	092B	64594
3	0033	61573
ळ	0933	59677
?	003F	55700
4	0034	51196
5	0035	45469
6	0036	41588
-	002D	40567
ओ	0913	39421
7	0037	38869
8	0038	37892
9	0039	35771
0	0030	35674
;	003B	33749
ढ	0922	14934
ई	0908	2988
+	002B	2758
ऊ	090A	2724
!	0021	1958
=	003D	1479
[	005B	1244
]	005D	1241
§	00A7	53
◌ॅ	0948	41
◌ः	0903	21
◌ौ	094C	16
'	0027	9
`	0060	6
ऐ	0910	5
औ	0914	4
^	005E	3
ऋ	090B	2
\	005C	1
:	003A	1

1. The result from character analysis with Devanagari sources.

1.1 The presence of Visarga (U+0903) in 4 files (totally 21 occurences):

abh01m.mul.xml
s0203m.mul.xml
s0504m.mul.xml
s0519m.mul.xml

The Visarga goes undetected in transformation; it exists even in the converted Roman sources. The character therefore has to be removed from the Roman sources. Also the validity of its presence in the Devanagari sources should be verified.

1.2 The presence of independent ai ऐ (U+0910) in 3 files (totally 5 occurences)

e0807n.nrf.xml
e1201n.nrf.xml
e1212n.nrf.xml

This character goes undetected in the converted Roman sources. It should be converted to a suitable letter. (In PP3, ē is used.)

1.3 The presence of dependent ai (U+0948) in 8 files (totally 41 occurences)

e0810n.nrf.xml
e1201n.nrf.xml
e1202n.nrf.xml
e1205n.nrf.xml
e1208n.nrf.xml
e1210n.nrf.xml
e1212n.nrf.xml
s0515m.mul.xml

This character goes undetected in the converted Roman sources. It should be converted to a suitable letter. (In PP3, ē is used.)

1.4 The presence of independent au औ (U+0914) in 3 files (totally 4 occurences)

e0807n.nrf.xml
e1201n.nrf.xml
e1212n.nrf.xml

This character goes undetected in the converted Roman sources. It should be converted to a suitable letter. (In PP3, ō is used.)

1.5 The presence of dependent au (U+094C) in 7 files (totally 16 occurences)

e0501n.nrf.xml
e0702n.nrf.xml
e0804n.nrf.xml
e0810n.nrf.xml
e1201n.nrf.xml
e1202n.nrf.xml
e1210n.nrf.xml

This character goes undetected in the converted Roman sources. It should be converted to a suitable letter. (In PP3, ō is used.)

1.6 The presence of vocalic r ऋ (U+090B) in 2 files (totally 2 occurences)

s0519m.mul.xml
vin12t.nrf.xml

As shown in the converted Roman sources, this letter can be removed.

1.7 The presence of section marks (§) in 13 files (totally 53 occurences)

s0102m.mul.xml
s0201m.mul.xml
s0203m.mul.xml
s0301m.mul.xml
s0502m.mul.xml
s0510m1.mul.xml
s0510m2.mul.xml
s0513m.mul.xml
s0514m.mul.xml
s0518m.nrf.xml
vin02m1.mul.xml
vin02m3.mul.xml
vin02m4.mul.xml

This symbol represents no meaning whatever in both Devanagari and Roman sources. This should be replaced with a space.

1.8 The presence of plain single quotes (') in 3 files (totally 9 occurences)

e0809n.nrf.xml
e1102n.nrf.xml
s0513a2.att.xml

To conform to the overall format, this should be converted to ‘ (U+2018) or ’ (U+2019).

1.9 The presence of backquotes (`) in 3 files (totally 6 occurences)

s0102m.mul.xml
s0103m.mul.xml
s0303m.mul.xml

To conform to the overall format, this should be converted to ‘ (U+2018)

1.10 The presence of circumflex (^) in 2 files (totally 3 occurences)

s0305t.tik.xml
s0401a.att.xml

This symbol should be removed.

1.11 The presence of backslash (\) in 1 files (totally 1 occurence)

abh03t.tik.xml

As shown in the converted Roman sources, this symbol can be removed.

1.12 The presence of colon (:) in 1 files (totally 1 occurence)

vin12t.nrf.xml

As shown in the converted Roman sources, this symbol can be removed.

==========

2. The report from textual comparison and corrections against the old CST4 data.

For this part, see the report presented in https://bhaddacak.github.io/correport.