Coding of Bangla Characters for Optical Recognition

Alamgir Mohammed
Assistant Professor, Dept of CSE, SUST
Email: alamgir99@gmail.com
Web: www.apona-bd.com

Abstract:

Coding of characters of some language is mostly necessary for the purpose of storage and transmission. For optical recognition of printed characters some form of codes are necessary. For languages like Bangla such codes do not exist. This work studies characteristics of Bangla texts and develops a coding for Bangla characters. The coding is done with the requirements of an optical charracter recogition system in mind.

Introduction:

Coding is assigning some unique number for a symbol of the alphabet of some language. Requirement for coding comes from the need of storing information in less memory space, and eventually needing less bandwidth when transferring from one place to another via communication media. A separate coding for optical recognition purpose is sometimes necessary when the recognition has some chance of better performance when characters are coded in some particualr way. For example, most usedly characters are often coded with *greater* hamming distances to minimize the chance of making error. This is in contrast with the usual coding of smaller hamming distance of frequently used characters in storage and communicatiotn. If a character recognition system does not gain anything from particualr coding then any arbitrary coding scheme can be used.

Distinct coding is not generally needed for OCR systems for langaugaes based on Latin alphabet because characters of such languages do not change appearance when used in words of written texts (except aesthetic ligatures). This is not the case with Indic languages which are derived from ancient Brahmi. For languages like Bangla and others (Devnagri etc) coding is more than just assigning different number for different charactes in the alphabet. Aplhabet of such languages only consists of vowels and consonants. But in writen texts a vast number of conjuncts, modified forms are used. From grammatical point of view, conjuncts and modified forms of characters are just the same as the constituent characters but shown in different contexts. Coding for Bengali in India has been standardized in this way. UNICODE coding for Bangla has been done in the same way. In Bangladesh, no coding exists (BSTI has never been successful) and different software developers have developed their own fonts with different numeric assignment.

Bangla Characters:

Bangla characters are primarily divided into vowels and consonants. When modifies a consonant, a vowel character is replaced by its short form, known as *kaar*. In Bangla a consonant can also modify another consonant or vowel. In such case, a short form of the consonant is used, called *fala*. Two or more consonants can combine and form a conjunct, known as *songjukto barna*. Some fala, kaar when used with certain character(s) may form allographs.

Vowels and kaars:

There are eleven vowels. When goes with a consonant, a vowel takes a shorter form and becomes kaar. The first vowel does not have any short form and is implicitly assumed to be with all consonants (unless there is a halant after a consonant). The vowels ি, ে and ৈ go before the consonant they modify while া,ী,ু ,ূ,ৃ, go after. Though seemingly ে া is formed by ে and া, ৌ by ে and ৗ; these kaars are grammatically different symbols and surprisingly goes around the character they modify. Within the same line, আ is not অ + া.

From grammatical point of view and purpose of coding for storage and transfer, the eleven vowels are unique symbols and demand unique codes. The ten kaars **may or may not** have unique codes. UNICODE has distinct code points for vowels and kaars. Therefore, ে া and ৌ have unique codes. For an OCR, আ is not any different from অ + া, same is for ে-া and ৌ. Therefore, for an OCR ে া and ৌ need not have distinct codes. Though আ may have a code point (otherwise where the া would be coming from).

অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
	া	ি	ী	ু	ূ	ৃ	ে	ৈ	ে - া	ে - ৗ

Figure 1: Bangla vowels and kaars.

Consonants:

There are about forty nine consonants in Bangla. The number is variable because of the scholarly debate about inclusion of some characters which are either conjuncts or some special form. Some consonants behave like vowels and have short forms similar to kaar. These forms are known as fala. Only ব, র and য have falas; they are: ্ব , ্র and ্য. Halant or hasant (্) is a special character that signifies a suppression of vowel (অ) after a consonant. Technically that acts as glue between consonants in formation of conjuncts. All consonants have merit for a unique code points. But the falas and conjuncts are considered sequence of consonants connected by halant; so they may be stored as such. Appearances of conjuncts are often too different from the constituent characters, so special rendering algorithm is needed for proper display of running texts. Similarly, to an OCR each conjunct appears as a distinctive shape different from others. As OCR deals with recognizing each shape no matter how they form, each conjunct that is *identifiable* should have a code point. The meaning of *identifiable* should be taken seriously as it relates to the capability of an OCR. A single conjunct may look like a single character to someone, while it may appear as a collection of different characters to others. For example, শ্ব may either be considered as a single or two separate characters (শ + ্ব, when they are separable and identifiable). Allographs are considered as unique characters and assigned separate codes.

Identifiability:

How the characters are segmented from words is an OCR design issue. In Bangla, the matra line is a good guide for recognition, and looking at characters from vertical and horizontal directions can largely isolate a character from others. In Pathak, codes are assigned with these two points in mind.

Font Issues:

Fonts are glyphs of characters of an alphabet. A font may have more than one glyph for a given character. Also, it may have more information than just the glyphs of characters. For Western languages fonts are almost an implementation of a coding table with each character having a

unique code point or position. For Bangla the characters or glyphs defined in a font file do not always match the concept of coding. For example, in major (non-UNICODE) commercial Bangla fonts, a number of characters have been defined multiple times. This is done to allow a better look of composed texts. Moreover, some conjuncts are defined partially with the hope that they would add up with other parts of adjacent character to form a conjunct. No one except the font developers has little knowledge about which parts combine with what. Since an OCR has to deal with legacy texts, character definitions in commercial fonts are required. The author has tried to find the codes of Bangla characters in commercial Bangla fonts. Among the commecial products, fonts from Bijoy are popular, but suffer from version incompatibility. Fonts that came with Bijoy 99 are stable and assumed to be ad-hoc standard.

Code List:

Here is the list of all symbols that Pathak recognizes as characters. The UNICODE and Bijoy equivalents are also given.

Table 1: Pathak Codes for Bangla Characters

AP code	Character	Form	Example	UNICODE (Hex)	Bijoy99 (Hex)	Prob %
1.	!	!		21	21	0.00
2.	‘	‘		26	D4	0.00
3.	’	’		27	D5	0.01
4.	৳	৳		9F3	24	0.00
5.	%	%		25	25	0.00
6.	৷	৷		??	40	0.00
7.	(	(		28	28	0.41
8.	)	)		29	29	0.37
9.	*	*		2A	2A	0.06
10.	+	+		2B	2B	0.13
11.	,	,		2C	2C	0.81
12.	-	-		2D	2D	0.42
13.	/	/		2F	2F	0.09
14.	:	:		3A	3A	0.18
15.	;	;		3B	3B	0.01
16.	<	<		3C	3C	0.00
17.	>	>		3E	3E	0.00
18.	=	=		3D	3D	0.29
19.	?	?		3F	3F	0.05
20.	।	।		964	7C	1.61
21.	॥	॥		965	5C	0.00
22.	০	০		9E6	30	1.46
23.	১	১		9E7	31	1.85
24.	২	২		9E8	32	0.59

25.	৩	৩		9E9	33	0.26
26.	৪	৪		9EA	34	0.40
27.	৫	৫		9EB	35	0.17
28.	৬	৬		9EC	36	0.19
29.	৭	৭		9ED	37	0.12
30.	৮	৮		9EE	38	0.20
31.	৯	৯		9EF	39	0.11
32.	অ	অ	অপনা	985	41	1.48
33.	ই	ই	ইতালি	987	42	1.21
34.	ঈ	ঈ	ঈদ	988	43	0.12
35.	উ	উ	উৎস	989	44	1.06
36.	ঊ	ঊ	ঊষা	98A	45	0.04
37.	ঋ	ঋ	ঋণ	98B	46	0.07
38.	এ	এ	একা	98F	47	1.39
39.	ঐ	ঐ	ঐরাবত	990	48	0.02
40.	ও	ও	ওরাও	993	49	0.71
41.	ঔ	ঔ	ঔষধ	994	4A	0.01
42.	া	া		9BE	76	10.62
43.	ে	ে		9C7	87	9.25
44.	ি	ি		9BF	77	5.27
45.	ী	ী		9C0	78	1.11
46.	র+ী	ী	বর্তী	?? 9C0	78 A9	0.02
47.	ু	ু		9C1	79	1.22
48.	ূ	ূ		9C2	7E	0.26
49.	ৃ	ৃ		9C3	84	0.18
50.	ৈ	ৈ		9C8	89	0.16
51.	ৗ	ৗ		9D7	8A	0.04
52.	ং	ং		982	73	0.99
53.	ঃ	ঃ		983	74	0.14
54.	ঁ	ঁ		981	75	0.07
55.	্	্		9CD	26	0.07
56.	্য	্য		9CD, 9AF	A8	1.72
57.	ক	ক	কলা	995	4B	4.95
58.	ক+ক	ক্ক	টেক্কা	995, 9CD, 995	B0	0.00
59.	ক+ট	ক্ট	ডক্টর	995, 9CD, 99F	B1	0.05
60.	ক+ত	ক্ত	রিক্ত	995, 9CD, 9A4	B3	0.14
61.	ক+ত+র	ক্ত্র	বক্ত্র	995, 9CD, 9A4, 9CD, 9B0	B3 AB	0.00
62.	র+ক	র্ক	তর্ক	9B0, 9CD, 995	4B A9	0.03

63.	ক+ব	ক্ব	পক্ব	995, 9CD, 9AC	4B A1	0.00
64.	ক+ম	ক্ম	রুক্মিণী	995, 9CD, 9AE	B4	0.00
65.	ক+র	ক্র	ক্রয়	995, 9CD, 9B0	B5	0.09
66.	ক+ল	ক্ল	ক্লান্ত	995, 9CD, 9B2	4B AC	0.05
67.	ক+ষ	ক্ষ	ক্ষেত্র	995, 9CD, 9B7	B6	0.25
68.	ক+ষ+ন	ক্ষ্ণ	তীক্ষ্ণ	995, 9CD, 9B7, 9CD, 9A8	B6 E8	0.00
69.	ক+ষ+ম	ক্ষ্ম	সূক্ষ্ম	995, 9CD, 9B8, 9CD, 9AE	B2	0.00
70.	ক+স	ক্স	কক্সবাজার	995, 9CD, 9B8	B7	0.09
71.	খ	খ	খর	996	4C	0.71
72.	র+খ	র্খ	মুর্খ	9B0, 9CD, 996	4C A9	0.00
73.	খ+র	খ্র	খ্রিষ্টাব্দ	996, 9CD, 9B0	4C AA	0.00
74.	গ	গ	গণনা	997	4D	0.92
75.	গ+ু	গু	গুণ	997, 9C1	B8	0.15
76.	গ+দ	গ্দ	?	997, 9CD, 9A6	BA	0.00
77.	গ+ধ	গ্ধ	দগ্ধ	997, 9CD, 9A7	BB	0.00
78.	গ+ন	গ্ন	লগ্ন	997, 9CD, 9A8	4D 9C	0.01
79.	গ+ব	গ্ব	দিগ্বিজয়	997, 9CD, 9AC	4D A6	0.00
80.	গ+ম	গ্ম	বাগ্মিতা	997, 9CD, 9AE	4D A5	0.00
81.	র+গ	র্গ	মার্গ	9B0, 9CD, 997	4D A9	0.00
82.	গ+র	গ্র	গ্রহণ	997, 9CD, 9B0	4D D6	0.10
83.	গ+র+ু	গ্রু	গ্রুপ	997, 9CD, 9B0, 9C1	4D D6 93	0.02
84.	গ+ল	গ্ল	গ্লানি	997, 9CD, 982	4D AD	0.00
85.	ঘ	ঘ	ঘর	998	4E	0.10
86.	র+ঘ	র্ঘ	অর্ঘ	9B0, 9CD, 998	4E A9	0.02
87.	ঘ+র	ঘ্র	ব্যাঘ্র	998, 9CD, 9B0	4E D6	0.00
88.	ঘ+ন	ঘ্ন	কৃতঘ্ন	998, 9CD, 9A8	4E 9C	0.00
89.	ঙ	ঙ	ব্যাঙ	999	4F	0.04
90.	ঙ+ক	ঙ্ক	অঙ্ক	999, 9CD, 995	BC	0.00
91.	ঙ+খ	ঙ্খ	শঙ্খ	999, 9CD, 996	95 4C	0.01
92.	ঙ+গ	ঙ্গ	বঙ্গ	999, 9CD, 997	BD	0.01
93.	ঙ+ঘ	ঙ্ঘ	লঙ্ঘন	999, 9CD, 998	95 4E	0.00
94.	ঙ+ঘ+র	ঙ্ঘ্র	অঙ্ঘ্রি	999, 9CD, 995, 9CD, 9B0	95 4E D6	0.00
95.	ঙ+ম	ঙ্ম	বাঙ্ময়	999, 9CD, 9AE	95 67	0.00
96.	চ	চ	চর	99A	50	0.62
97.	চ+চ	চ্চ	উচ্চ	99A, 9CD 99A	94 50	0.02
98.	চ+ছ	চ্ছ	ইচ্ছা	99A, 9CD 99B	94 51	0.09
99.	চ+ছ+ব	চ্ছ্ব	উচ্ছ্বাস	99A, 9CD 99B, 9CD, 9AC	94 51 A1	0.00
100.	চ+ছ+র	চ্ছ্র	উচ্ছ্রয়	99A, 9CD 99B, 9CD, 9BO	94 51 AB	0.00

101.	র+চ	র্চ	চর্চা	9B0, 9CD 99A	50 A9	0.00
102.	ছ	ছ	ছবি	99B	51	0.46
103.	জ	জ	জড়তা	99C	52	0.93
104.	জ+জ	জ্জ	লজ্জা	99C, 9CD, 99C	BE	0.00
105.	জ+জ+ব	জ্জ্ব	উজ্জ্বল	99C, 9CD, 99C, 9CD, 9AC	BE A1	0.00
106.	জ+ঝ	জ্ঝ	কুজ্ঝটিকা	99C, 9CD, 99D	C0	0.00
107.	জ+ঞ	জ্ঞ	জ্ঞান	99C, 9CD, 99E	C1	0.02
108.	জ+ব	জ্ব	জ্বলন	99C, 9CD, 9AC	52 A1	0.01
109.	র+জ	র্জ	বর্জ	9B0, 9CD, 99C	52 A9	0.00
110.	জ+র	জ্র	বজ্র	99C, 9CD, 9B0	52 AA	0.00
111.	ঝ	ঝ	ঝরা	99D	53	0.15
112.	ঞ	ঞ	মিঞা	99E	54	0.01
113.	ঞ+চ	ঞ্চ	অঞ্চল	99E, 9CD, 99A	C2	0.01
114.	ঞ+ছ	ঞ্ছ	বাঞ্ছা	99E, 9CD, 99B	C3	0.00
115.	ঞ+জ	ঞ্জ	অঞ্জলি	99E, 9CD, 99C	C4	0.01
116.	ঞ+ঝ	ঞ্ঝ	ঝঞ্ঝা	99E, 9CD, 99D	C5	0.00
117.	ট	ট	টক	99F	55	2.17
118.	ট+ট	ট্ট	চট্টগ্রাম	99F, 9CD, 99F	C6	0.00
119.	ট+ব	ট্ব	খট্বাঙ্গ	99F, 9CD, 9AC	55 A1	0.00
120.	ট+ম	ট্ম	কুট্মল	99F, 9CD, 9AE	55 A5	0.00
121.	র+ট	র্ট	মর্টার	9B0. 9CD, 99F	55 A9	0.00
122.	ট+র	ট্র	ট্রাক	99F, 9CD, 9B0	55 AA	0.02
123.	ঠ	ঠ	ঠাকুর	9A0	56	0.09
124.	ড	ড	ডাব	9A1	57	0.44
125.	ড+ড	ড্ড	আড্ডা	9A1, 9CD, 9A1	C7	0.00
126.	ড+র	ড্র	ড্রাম	9A1, 9CD, 9B0	57 AA	0.00
127.	র+ড	র্ড	জর্ডান	9B0, 9CD, 9A1	57 A9	0.03
128.	ঢ	ঢ	ঢাকা	9A2	58	0.04
129.	ঢ+ম	ঢ্ম	অঢ্ম	9A2, 9CD, 9AE	58 A5	0.00
130.	ঢ+র	ঢ্র	মেঢ্রো	9A2, 9CD, 9B0	58 AA	0.00
131.	ণ	ণ	গুণ	9A3	59	0.48
132.	ণ+ট	ণ্ট	ঘণ্টা	9A3, 9CD, 99F	C8	0.00
133.	ণ+ঠ	ণ্ঠ	কণ্ঠ	9A3, 9CD, 9A0	C9	0.00
134.	ণ+ড	ণ্ড	পণ্ডিত	9A3, 9CD, 9A1	CA	0.00
135.	ণ+ণ	ণ্ণ	বিষণ্ণ	9A3, 9CD, 9A3	59 9C	0.00
136.	ণ+ব	ণ্ব	কণ্ব	9A3, 9CD, 9AC	59 5E	0.00
137.	ণ+ম	ণ্ম	হিরণ্ময়	9A3, 9CD, 9AE	59 A5	0.00
138.	র+ণ	র্ণ	বর্ণ	9B0, 9CD, 9A3	59 A9	0.07

139.	ত	ত		9A4	5A	2.56
140.	ত+ত	ত্ত	সত্তা	9A4, 9CD, 9A4	CB	0.08
141.	ত+ত+ব	ত্ত্ব	সত্ত্ব	9A4, 9CD, 9A4, 9CD, 9AC	CB A1	0.00
142.	ত+থ	ত্থ	উত্থান	9A4, 9CD, 9A5	CC	0.00
143.	ত+ন	ত্ন	রত্ন	9A4, 9CD, 9A8	5A 9C	0.00
144.	ত+ব	ত্ব	ত্বক/সত্বর	9A4, 9CD, 9AC	5A A1	0.02
145.	ত+ম	ত্ম	আত্মা	9A4, 9CD, 9AE	5A A5	0.02
146.	ত+র	ত্র	ত্রপা/পত্র	9A4, 9CD, 9B0	CE	0.26
147.	ত+র+ু	ত্রু	ত্রুটি	9A4, 9CD, 9B0, 9C1	CE 93	0.00
148.	র+ত	র্ত	আর্ত	9B0, 9CD, 9A4	5A A9	0.26
149.	থ	থ		9A5	5F	0.51
150.	থ+ব	থ্ব	পৃথ্বি	9A5, 9CD, 9AC	5F A1	0.00
151.	র+থ	র্থ	অর্থ	9A5, 9CD, 9A5	5F A9	0.11
152.	থ+র	থ্র	থ্র	9A5, 9CD, 9B0	5F AA	0.00
153.	থ+র+ু	থ্রু		9A5, 9CD, 9B0, 9C1	5F AA 93	0.00
154.	দ	দ		9A6	60	1.28
155.	দ+গ	দ্গ	উদ্গাত	9A6, 9CD, 997	98 4D	0.00
156.	দ+ঘ	দ্ঘ	উদ্ঘাটন	9A6, 9CD, 998	98 4E	0.00
157.	দ+দ	দ্দ	উদ্দাম	9A6, 9CD, 9A6	CF	0.00
158.	দ+দ+ব	দ্দ্ব	এতদ্দ্বারা	9A6, 9CD, 9A6, 9CD, 9AC	CF A1	0.00
159.	দ+ধ	দ্ধ	রুদ্ধ	9A6, 9CD, 9A7	D7	0.19
160.	দ+ব	দ্ব	বিদ্বান	9A6, 9CD, 9AC	D8	0.05
161.	দ+ভ	দ্ভ	সদ্ভাব	9A6, 9CD, 9AD	99 A2	0.00
162.	দ+ম	দ্ম	পদ্ম	9A6, 9CD, 9AE	D9	0.00
163.	দ+র	দ্র	দ্রবণ	9A6, 9CD, 9B0	60 AA	0.01
164.	দ+র+ু	দ্রু	দ্রুত	9A6, 9CD, 9B0. 9C1	60 AA 93	0.02
165.	র+দ	র্দ	গর্দান	9B0, 9CD, 9A6	60 A9	0.07
166.	র+দ+র	র্দ্র	আর্দ্রতা	9B0, 9CD, 9AC, 9CD, 9B0	60 AA A9	0.00
167.	ধ	ধ	ধন	9A7	61	0.59
168.	ধ+ন	ধ্ন	গৃধ্নু	9A7, 9CD, 9A8	61 9C	0.00
169.	ধ+ব	ধ্ব	ধ্বনি	9A7, 9CD, 9AC	61 9F	0.00
170.	ধ+ম	ধ্ম	অধ্মান	9A7, 9CD, 9AE	61 A5	0.00
171.	ধ+র	ধ্র	ধ্রব	9A7, 9CD, 9B0	61 AA	0.00
172.	ধ+র+ু	ধ্রু	ধ্রুব	9A7, 9CD, 9B0, 9C1	61 AA 93	0.00
173.	র+ধ	র্ধ	অর্ধ	9B0, 9CD, 9A7	61 A9	0.03
174.	ন	ন	নাক	9A8	62	3.93
175.	ন+ট	ন্ট	টেন্ট	9A8, 9CD, 99F	9B 55	0.07
176.	ন+ঠ	ন্ঠ	লুন্ঠন	9A8, 9CD, 9A0	DA	0.00

177.	ন+ড	ন্ড	গুন্ডা	9A8, 9CD, 9A1	DB	0.12
178.	ন+ত	ন্ত	অন্ত	9A8, 9CD, 9A4	9A 97	0.15
179.	ন+ত+ব	ন্ত্ব	সান্ত্বনা	9A8, 9CD, 9A4, 9CD, 9AC	9A 97 A1	0.00
180.	ন+ত+র	ন্ত্র	যন্ত্র	9A8, 9CD, 9A4, 9CD, 9B0	9A BF	0.07
181.	ন+থ	ন্থ	গ্রন্থ	9A8, 9CD, 9A5	9A 92	0.00
182.	ন+দ	ন্দ	আনন্দ	9A8, 9CD, 9A6	9B 60	0.02
183.	ন+দ+ব	ন্দ্ব	দ্বন্দ্ব	9A8, 9CD, 9A6, 9CD, 9AC	9B D8	0.00
184.	ন+দ+র	ন্দ্র	চন্দ্র	9A8, 9CD, 9A6, 9CD, 9B0	9B 60 AA	0.00
185.	ন+ধ	ন্ধ	বন্ধ	9A8, 9CD, 9A7	DC	0.01
186.	ন+ধ+র	ন্ধ্র	অন্ধ্র	9A8, 9CD, 9A7, 9CD, 9B0	DC AB	0.00
187.	ন+ন	ন্ন	অন্ন	9A8, 9CD, 9A8	62 9C	0.07
188.	ন+ব	ন্ব	অন্বিত	9A8, 9CD, 9AC	9A 5E	0.02
189.	ন+ম	ন্ম	জন্ম	9A8, 9CD, 9AE	62 A5	0.02
190.	প	প	পথ	9AA	63	2.19
191.	প+ট	প্ট	চ্যাপ্টার	9AA, 9CD, 99F	DE	0.00
192.	প+ত	প্ত	সপ্ত	9AA, 9CD, 9A4	DF	0.03
193.	প+ন	প্ন	স্বপ্ন	9AA, 9CD, 9A8	63 9C	0.00
194.	প+প	প্প	গপ্প	9AA, 9CD, 9AA	E0	0.00
195.	প+র	প্র	প্রথম	9AA, 9CD, 9B0	63 D6	0.79
196.	প+র+ু	প্রু	প্রুফ	9AA, 9CD, 9B0, 9C1	63 D6 93	0.00
197.	র+প	র্প	অর্পণ	9B0, 9CD, 9AA	63 A9	0.00
198.	প+ল	প্ল	প্লাবন	9AA, 9CD, 9B2	63 AD	0.02
199.	প+স	প্স	অপ্সরা	9AA, 9CD, 9B8	E1	0.00
200.	ফ	ফ	ফুল	9AD	64	0.45
201.	ফ+র	ফ্র	আফ্রিকা	9AD, 9CD, 9B0	64 AB	0.01
202.	র+ফ	র্ফ	সার্ফ	9B0, 9CD, 9AD	64 A9	0.00
203.	ফ+ল	ফ্ল	ফ্লপি	9AD, 9CD, 9B2	64 AC	0.16
204.	ব	ব	বক	9AC	65	3.91
205.	ব+জ	ব্জ	কব্জা	9AC, 9CD, 99C	E2	0.00
206.	ব+দ	ব্দ	শব্দ	9AC, 9CD, 9A6	E3	0.01
207.	ব+ধ	ব্ধ	লব্ধ	9AC, 9CD, 9A7	E4	0.00
208.	ব+ব	ব্ব	আব্বা	9AC, 9CD, 9AC	65 9F	0.00
209.	ব+র	ব্র	ব্রত	9AC, 9CD, 9B0	65 AA	0.00
210.	ব+র+ু	ব্রু	ব্রুনাই	9CA, 9CD, 9B0, 9C1	65 AA 93	0.00
211.	র+ব	র্ব	গর্ব	9B0, 9CD, 9AC	65 A9	0.07
212.	ব+ল	ব্ল	ব্লক	9AC, 9CD, 9B2	65 AD	0.02
213.	ভ	ভ	ভাল	9AD	66	0.38
214.	ভ+র	ভ্র	ভ্রম	9AD, 9CD, 9B0	E5	0.00

215.	র+ভ	র্ভ	প্রাদুর্ভাব	9B0, 9CD, 9AD	66 A9	0.05
216.	ভ+র+ু	ভ্রু	ভ্রুকুটি	9B0, 9CD, 9AD, 9C1	E5 93	0.00
217.	ভ+র+ূ	ভ্রূ	ভ্রূণ	9B0, 9CD, 9AD, 9C2	E5 83	0.00
218.	ম	ম	মা	9AE	67	1.97
219.	ম	ম	কম্প	9AE	A4	0.00
220.	ম+ন	ম্ন	নিম্ন	9AE, 9CD, 9A8	67 9C	0.01
221.	ম+প	ম্প	কম্প	9AE, 9CD, 9AA	A4 FA	0.34
222.	ম+প+র	ম্প্র	কম্প্র	9AE, 9CD, 9AA, 9CD, 9B0	A4 63 D6	0.00
223.	ম+ফ	ম্ফ	লম্ফ	9AE, 9CD, 9AB	E7	0.00
224.	ম+ব	ম্ব	লম্ব	9AE, 9CD, 9AC	A4 5E	0.01
225.	ম+ভ	ম্ভ	দম্ভ	9AE, 9CD, 9AD	A4 A2	0.06
226.	ম+ভ+র	ম্ভ্র	সম্ভ্রম	9AE, 9CD, 9AD, 9CD, 9B0	A4 A3	0.00
227.	ম+ম	ম্ম	সম্মান	9AE, 9CD, 9AE	A4 A7	0.01
228.	ম+র	ম্র	ম্রিয়মাণ	9AE, 9CD, 9B0	67 AA	0.00
229.	র+ম	র্ম	মর্মর	9B0, 9CD, 9AE	67 A9	0.06
230.	ম+ল	ম্ল	ম্লান	9AE, 9CD, 9B2	A4 AD	0.00
231.	য	য	যান	9AF	68	1.01
232.	র+য	র্য	আর্য	9B0, 9CD, 9AF	68 A9	0.06
233.	য়	য়	আয়	9DF	71	2.05
234.	র	র	রক	9B0	69	6.52
235.	র+ু	রু	রুপা	9B0, 9C1	69 93	0.08
236.	র +ূ	রূ	রূঢ়	9B0, 9C2	69 83	0.00
237.	ল	ল	লাল	9B2	6A	2.22
238.	ল+ক	ল্ক	শুল্ক	9B2, 9CD, 995	E9	0.00
239.	ল+গ	ল্গ	বল্গা	9B2, 9CD, 997	EA	0.00
240.	ল+ট	ল্ট	বেল্ট	9B2, 9CD, 99F	EB	0.02
241.	ল+ড	ল্ড	গোল্ড	9B2, 9CD, 9A1	EC	0.00
242.	ল+প	ল্প	অল্প	9B2, 9CD, 9AA	ED	0.02
243.	ল+ফ	ল্ফ	উল্ফন	9B2, 9CD, 9AB	EE	0.00
244.	ল+ব	ল্ব	বিল্ব	9B2, 9CD, 9AC	6A A6	0.00
245.	ল+ম	ল্ম	গুল্ম	9B2, 9CD, 9AE	6A A5	0.00
246.	ল+ল	ল্ল	বল্লরি	9B2, 9CD, 9B2	6A AD	0.01
247.	র+ল	র্ল	পার্ল	9B0, 9CD, 9B2	6A A9	0.00
248.	শ	শ	শাক	9B6	6B	0.85
249.	শ +ু	শু	শুল	9B6, 9C1	EF	0.07
250.	শ+চ	শ্চ	নিশ্চিত	9B6, 9CD, 99A	F0	0.00
251.	শ+ছ	শ্ছ	নিশ্ছিদ্র	9B6, 9CD, 99B	F1	0.00
252.	শ+ন	শ্ন	প্রশ্ন	9B6, 9CD, 9A8	6B 9C	0.01

253.	শ+ব	শ্ব	শ্বশুর	9B6, 9CD, 9AC	6B A6	0.02
254.	শ+ম	শ্ম	শ্মশান	9B6, 9CD, 9AE	6B A5	0.00
255.	শ+র	শ্র	শ্রাবণ	9B6, 9CD, 9B0	6B AA	0.00
256.	শ+র	শ্রু	শ্রুতি	9B6, 9CD, 9B0, 9C1	6B D6 93	0.00
257.	শ+ল	শ্ল	শ্লীল	9B6, 9CD, 9B2	6B AD	0.01
258.	ষ	ষ		9B7	6C	0.22
259.	ষ	ষ	বাষ্প	9B7	AE	0.00
260.	ষ+ক	ষ্ক	শুষ্ক	9B7, 9CD, 995	AE 8B	0.00
261.	ষ+ক+র	ষ্ক্র	নিষ্ক্রমণ	9B7, 9CD, 995, 9CD, 9B0	AE 8C	0.00
262.	ষ+ট	ষ্ট	শিষ্ট	9B7, 9CD, 99F	F3	0.11
263.	ষ+ট+র	ষ্ট্র	রাষ্ট্র	9B7, 9CD, 99F, 9CD, 9B0	F3 AA	0.00
264.	ষ+ঠ	ষ্ঠ	জৈষ্ঠ	9B7, 9CD, 9A0	F4	0.02
265.	ষ+ণ	ষ্ণ	তৃষ্ণা	9B7, 9CD, 9A3	F2	0.00
266.	ষ+প	ষ্প	পুষ্প	9B7, 9CD, 9AA	AE 63	0.00
267.	ষ+ফ	ষ্ফ	নিষ্ফলা	9B7, 9CD, 9AB	F5	0.00
268.	ষ+ব	ষ্ব	পিতৃষ্বসা	9B7, 9CD, 9AC	6C A6	0.00
269.	ষ+ম	ষ্ম	গ্রীষ্ম	9B7, 9CD, 9B7	AE A7	0.00
270.	র+ষ	র্ষ	বিমর্ষ	9B0, 9CD, 9B7	6C A9	0.03
271.	স	স	সাপ	9B8	6D	2.36
272.	স	স	স্পর্শ	9B8	AF	0.00
273.	স+ক	স্ক	পুরস্কার	9B8, 9CD, 995	AF 8B	0.02
274.	স+ক+র	স্ক্র	স্ক্রু	9B8, 9CD, 995, 9CD, 9B0	AF 8C	0.00
275.	স+খ	স্খ	স্খলন	9B8, 9CD, 996	F6	0.00
276.	স+ট	স্ট	মাস্টার	9B8, 9CD, 99F	F7	0.08
277.	স+ট+র	স্ট্র	স্ট্রিট	9B8, 9CD, 99F, 9CD, 9B0	F7 AA	0.00
278.	র+স+ট	র্স্ট	ফার্স্ট	9B0, 9CD, 9B8, 9CD, 99F	F7 A9	0.00
279.	স+ত	স্ত	অস্ত	9B8, 9CD, 9A4	AF 97	0.10
280.	স+থ	স্থ	স্থান	9B8, 9CD, 9A5	AF 92	0.19
281.	স+ন	স্ন	স্নান	9B8, 9CD, 9A8	6D 9C	0.00
282.	স+প	স্প	স্পর্শ	9B8, 9CD, 9AA	AF FA	0.00
283.	স+ফ	স্ফ	স্ফোটন	9B8, 9CD, 9AB	F9	0.00
284.	স+ব	স্ব	স্বামী	9B8, 9CD, 9AC	AF 5E	0.03
285.	স+ম	স্ম	স্মরণ	9B8, 9CD, 9AE	AF A7	0.03
286.	স+র	স্র	স্রোত	9B8, 9CD, 9B0	6D AA	0.00
287.	র+স	র্স	পার্স	9B0, 9CD, 9B8	6D A9	0.01
288.	স+ল	স্ল	স্লেট	9B8, 9CD, 9B2	AF AD	0.00
289.	হ	হ	হাত	9B9	6E	1.74
290.	হ+উ	হু	হুশ	9B9, 9C1	FB	0.02

291.	হ+ঋ	হৃ	হৃদয়	9B9, 9C3	FC	0.06
292.	হ+ণ	হ্ণ	অপরাহ্ণ	9B9, 9CD, 9A3	6E E8	0.00
293.	হ+ন	হ্ন	মধ্যাহ্ন	9B9, 9CD, 9A8	FD	0.03
294.	হ+ব	হ্ব	আহ্বান	9B9, 9CD, 9AC	6E 9F	0.00
295.	হ+ম	হ্ম	ব্রহ্মান্ড	9B9, 9CD, 9AE	FE	0.00
296.	হ+র	হ্র	হ্রাস	9B9, 9CD, 9B0	6E AB	0.00
297.	হ+ল	হ্ল	আহ্লাদ	9B9, 9CD, 9B2	6E AC	0.00
298.	ড়	ড়	বাড়তি	9DC	6F	0.21
299.	ঢ়	ঢ়	?	9DD	70	0.03
300.	ৎ	ৎ	উৎপাত	9fc (v4.1)	72	0.21

Frequency Information:

Though frequency of different vowels and consonants has been reported in literature, the information about conjuncts is yet unavailable. The main reason for such unavailability is the wide variation in commercial fonts used in composing sample documents. The author has attempted to extract some frequency information of conjuncts from sample documents written in Bijoy fonts. To be perfect, the study acknowledges each of the Pathak symbol defined in the code table as a unique symbol and compiles references of equivalent Bijoy code string. Details of the frequency distribution study are given next. This frequency information is used in the recognition process of Pathak.

Any presence of a consonant in sample data does not necessarily mean a frequency count for that character. Rather, if any other character is being with that in formation of a consonant is also considered. Since in the code list, single consonants come before their conjuncts, to make the work easier the code table is reversed. That means conjuncts of a consonant are looked before the consonant itself. The comparison is between strings of variable length, so, conjuncts with the highest number of constituents are moved top in the list. To be specific, 3-character conjuncts are in the top followed by 2-character and then the isolated ones.

The comparison is done by first taking three bytes from the sample document and looking for if any of the 3-character conjunct matches. If a match is found, new search begins. Otherwise, first two bytes are matched against any of 2-character conjuncts. If any match is found a new search begins with the leftover returned. On fail, the first byte is looked for the single character match.

Before any search begins in the sample texts, multiple codes of single character are replaced with the first code. For example, ু kaar has three codes: 79, 7A, 96; any occurences of 7A and 96 are replaced by 79.

Frequency results are given in last column of Table 1. Please note, the corpus used for the frequency distribution study was very limited, about a million words, so the probabilties are not much reliable. The author is trying to have access to large electronic files written in Bijoy.