

PuthuNira – A Set of New Keyboard Layouts for Malayalam

Ajith R. MBBS, DCP

Nirvriti, T.C. 5/3457(3), Kezhuvankara

Pongumoodu, Medical College P.O.

Thiruvananthapuram, Kerala 695 011

ajith@disroot.org

+91 89390 01898
=====

1.1. Abstract

The conventional keyboard layout schemes used by Malayalam, one of the Indian languages, do not consider the ease of typing the language. This study suggests two new keyboard layouts for Malayalam designed using character usage frequency, describes methods for comparing ease of typing and uses the methods to compare the two commonly used keyboard layouts InScript and Remington with the newly suggested layouts PuthuNira Aarambham and Puthunira Charutha. The study result suggests that the new layouts will be easier to type than conventional Malayalam keyboard layout.

1.2. Introduction

Malayalam is the language native to the people of Kerala, a state in south India and spoken by over 38 million (Gutman and Avanzati 2013). Malayalam has an abugida system of writing (Gutman and Avanzati 2013). The number of glyphs originally required to write Malayalam was over 500. This was brought down to 90 following an orthography reform in 1967 with further modifications in 1969 (Government of Kerala 1971). With the advent of computing, Malayalam, along with other Indian languages, was represented using ISCII (“INDIAN SCRIPT CODE FOR INFORMATION INTERCHANGE - ISCII” 1991). ISCII has been rendered largely obsolete with the rising adoption of Unicode to represent Malayalam (Kerala Government Information Technology Department 2008). Unicode 15 represents 118 Malayalam characters, of which 69 are regularly used (“The Unicode® Standard Version 15.0 – Core Specification” 2022). Unicode doesn’t have separate encodings for consonant signs. In conjuncts, zero width non joiner is used to display the consonant form rather than its sign.

Malayalam computer keyboard layout followed the Remington typewriter sequence initially (“Remington” 2010). Ralminov and others have modified Remington layouts modifying the original by adding the missing keys (Ralminov 2008), (“ThoolikaUnicode”, 2006). Phonetic transliteration schemes used to type Malayalam include Mozhi, Swanalekha, Google and others.

InScript layout standardized by the Government of India and approved by the Government of Kerala is the decreed standard Malayalam keyboard layout, which was enhanced recently (“A Document for Enhanced InScript Keyboard Layout 5.2” 2010).

Remington keyboard layout was adapted for computers probably because it was the main keyboard layout used in typewriters. The original keyboard layout was designed by omitting certain characters to squeeze the most important characters into the limited number of keys in a typewriter (Nair 1971). When the typewriter keyboard was adapted for computers, the principles by which the typewriter keyboard was designed were not considered. Instead, the missing characters were placed into the top row replacing the symbols in that row.

InScript layout is designed to keep the mapping of the characters of all Indian languages common (phonetic design) thereby making it advantageous for a person who knows typing in one Indian script to type in any other Indian script. InScript claims to simplify the layout by placing all the vowels on the left side of the keyboard layout and the consonants on the right side, taking advantage of the division of the Indian languages (“A Document for Enhanced InScript Keyboard Layout 5.2” 2010). However, this will only make the layout easier to remember, not necessarily easier to type. The phonetic transliteration schemes rely on typing an English phrase phonetically similar to the Malayalam word(s) required. We can easily appreciate that none of the keyboard layouts have taken into account the ease of typing.

There aren't any universally accepted methods or metrics to assess ease of typing of a keyboard layout. This situation arises because ease of typing is a poorly measurable characteristic that is influenced by personal factors like hand preference, absence of fingers, the type of matter that is typed frequently etc. Within these constraints, the more common metrics that are used to assess keyboard layouts are the distance traveled by each finger, frequency of consecutively using the same finger, change of fingers in sequence (rolling) either inwards or outwards versus alternating hands versus repeated use of same finger when striking consecutive characters etc (Krzywinski, n.d.), (Capewell, n.d.), (Gillespie, n.d.), (Bucao 2010). These metrics were mainly used in assessing keyboard layouts of the English language.

When assessing languages like Malayalam where the number of basic characters far exceeds that of English, we should also consider the total number of key presses for a given passage and the effect of modifier keys.

1.3 Objectives

1. To ascertain from a representative Malayalam passage
 - + the frequency of each character
 - + the frequency of characters following it and their frequency in that position.
2. To modify the representative Malayalam passage to change the characters such that it would

have the same character for vowels and their signs (with the exception of chandrakala (ഞ) being mapped to അ), gemination of consonants would be effected by typing a consonant followed by a special key, vowel less forms (chillu) are formed by typing a consonant followed by another special key and to ascertain from it

- + the frequency of each character
- + the frequency of characters following it and their frequency in that position.

3. To develop, using the frequency data, two Malayalam keyboard layouts that would make typing Malayalam easier. One keyboard with statically assigned keys and one such that the changes in (2) above are possible while typing.

4. To compare the so developed Malayalam keyboard layouts with the existing Remington and InScript layouts.

1.4 Materials and Methods

1.4.1 Estimating Character Frequencies

The SMC Malayalam corpus was downloaded and combined into a single file (“Malayalam Corpus by Swathanthra Malayalam Computing” 2019). HTML tags, characters other than Malayalam except zero width joiner and zero width non-joiner were removed from the text. Chillu characters represented traditionally were converted into atomic representation. The various representations of conjunct റ്റ were converted into an uniform representation റ്റ. Similarly, റ was converted to the phonetically correct representation of റ്റ. Spaces were merged and converted to open boxes. These changes were made using a Perl script named ‘changesCommon.pl’ and is available in the code repository for the project (link given in End Notes). It was randomly split into a training and a testing set in the ratio 60:40.

The training set was further modified to facilitate objectives one and two. One set of changes effected by using ‘changesAarambham.sed’ was to convert the consonant symbols to correct for the restrictions imposed by fonts. Malayalam fonts in current use impose that the consonant symbol ഞ is allowed only if you type അ, but not റ, though language rules are clear that the consonant symbol is common for both consonants. Similarly, the symbol ഞ is common for റ and റ, though fonts permit it only for റ. The correction followed the principles given in (Ramakumar V 2004). Frequencies of each character, the characters that follow, their frequencies, proportions and ranks were calculated using standard Linux utilities grep, uniq, sort, paste, sed and awk. The script files used were named ‘RPcalc.sh’ and ‘RPpaircalc.sh’ and are available along with the commands used in the code repository for the project.

The second set of changes were made before ascertaining frequency to fulfill objective 2. The changes were effected using ‘changesCharutha.sed’ and included the changes mentioned in the para above and more. These included converting geminated consonants into base character

plus a special character :, converting chillus to a base character plus a special character ⊖, converting all vowel symbols to their corresponding vowel, converting chandrakala ് to അ, converting ണ് to ണ, converting ് to ള. These changes were made to ascertain the character frequency for developing a keyboard that would dynamically assign the same key to vowels or their vowel signs as dictated by the cursor position in a word, that would geminate a consonant by pressing a specially assigned key, that would convert a consonant to chillu if another specially assigned key would be pressed after the consonant and in which the conjuncts ണ് and ള would be assigned to their unique key positions rather than to the sequence of their constituent consonants.

1.4.2 Designing New Keyboard Layouts

To design a new Keyboard layout, a hypothetical, idealised, (8+8) * 5 + (1 + 1) grid of keys was assumed for simplicity (Figure 1). A second grid of the same dimensions was also considered for the ‘Shift’ (or another modifier) positions of the first grid. Each key in these grids was given a rank denoting the preference for the key. Any layout may have one or more of these 82+82 ranked positions unassigned. While evolving an unbiased ranking is considered not feasible, certain principles were adhered to in ranking the keys:

1. Home row keys were assigned higher ranks (lower numerals).
2. Stronger fingers were assigned higher ranks with the ranked order of fingers being thumb, index, middle, ring and little.
3. Keys that were to be pressed after lifting fingers from their home positions were ranked giving preference to straight up movements followed by straight down, followed by horizontal followed by oblique up and finally oblique down directions from the home positions for all fingers except the little finger. For the little finger, downward movements (straight and oblique) were given higher ranks over corresponding upward movements.
4. Grid positions that necessitate movements were ranked according to the distance to be traversed from the home position, with lower ranks assigned to greater finger travel.
5. For positions that were away by more than a row or column from home position, ranks assigned were lower than those assigned to ‘Shift’ed keys in positions within one key of the home positions.
6. Both hands were considered equal.

The grids with the ranks assigned to each key and the English characters that will fall in each of these keys in a standard QWERTY keyboard layout is given in figure 1. There are two panes, one for the ‘unshifted’ and another for the ‘shifted’ positions. In the figure, for those positions not available in a standard 105 keyboard, only ranks are shown. The positions that are not more than one key away from home positions are bounded by a dotted rectangle. The y-axis label is 0 for the home row and increasingly negative downwards and increasingly positive upwards. The “L” in x-axis labels stand for left and “R” for right. A vertical black line separates

the left and right half of the grid. Columns are numbered from 0 to 8 to either side. The column 0 indicates home column of thumb (usually the space bar), 2 index finger's home, 1 index finger moved inwards one column, 3 to 5 home columns of middle, ring, and little finger respectively and 6 to 8 columns accessed by the little finger by moving outwards by increasing number of columns.

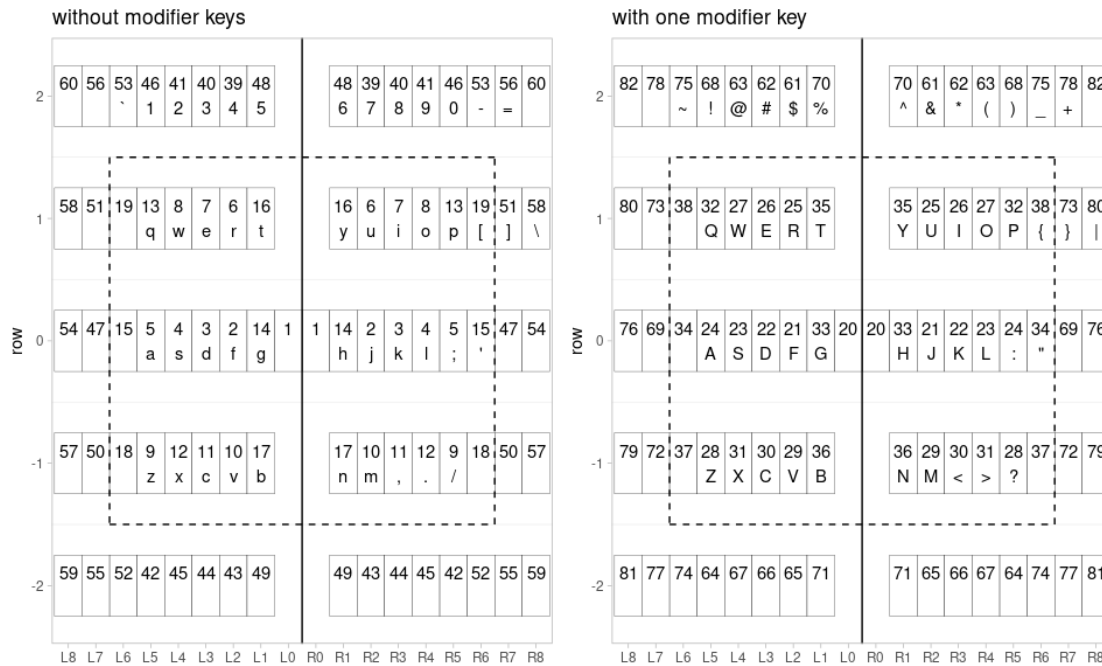


Figure 1: Rank for each key in an idealised matrix and corresponding characters in QWERTY layout

InScript and Remington keyboard layouts are shown in figure 2 to show the position of each character. The empty positions are not shown. The characters in ‘shift’ position are shown above those in ‘unshift’ position in each cell. Ranks of each position are not repeated. X and y axes are marked as in figure 1. Positions that fall within one key away from home positions are shown bounded with dotted lines.

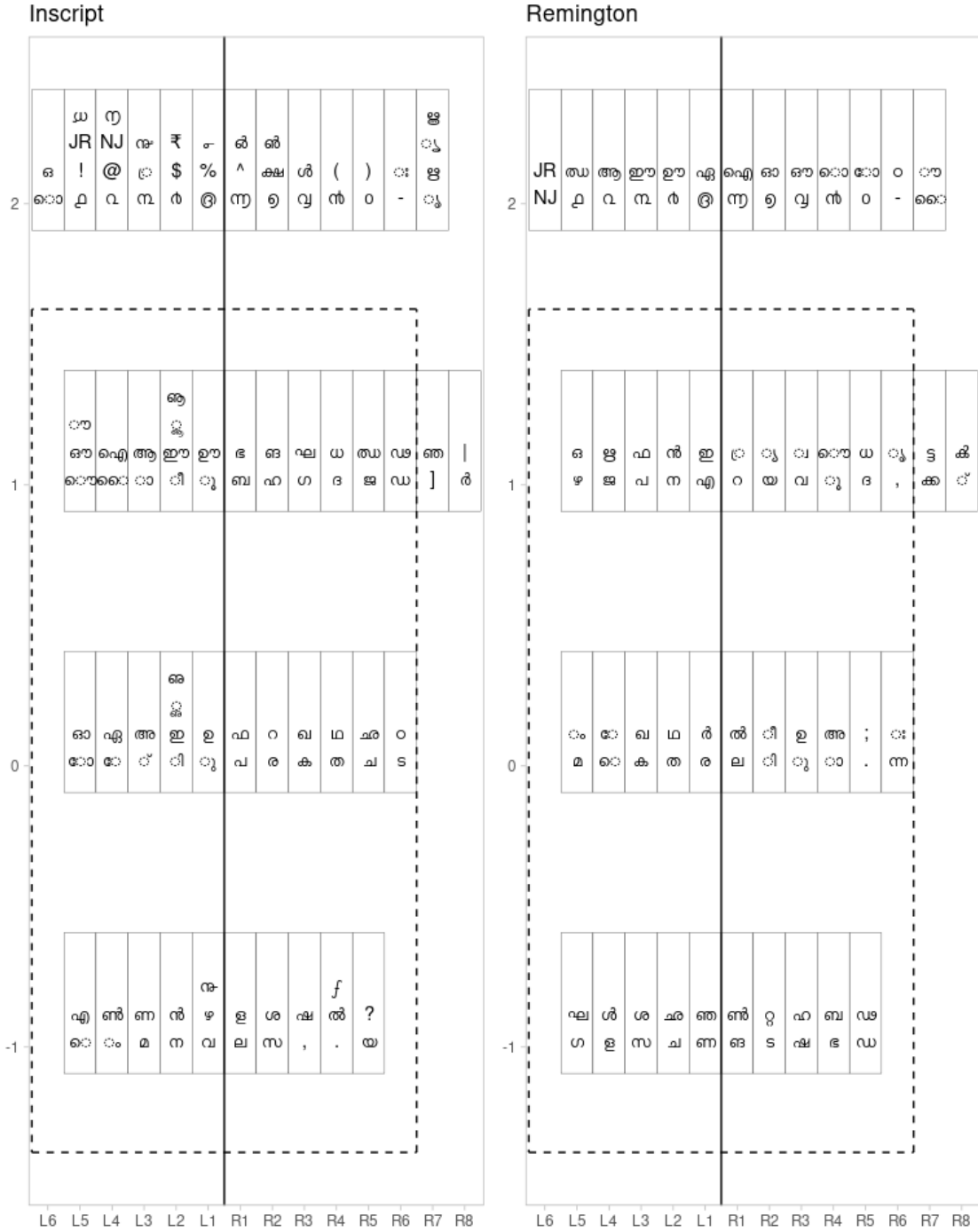


Figure 2: Inscript and Remington layouts

Characters were ranked according to their frequency in the analysed Malayalam text. Consecutive pairs were assigned to key positions of the highest rank (lowest numeral) key positions of either side. The next pair was assigned to the next highest rank, but the sides were alternated. Thus, frequency ranks 1,4,5 would fall on keys ranked 1,2,3 on one side while frequency ranks 2,3,6 would fall on the same ranked keys on the other side. However, perfect

matching was not attempted. Rank matching was relaxed to change the hand or finger if a character's most frequent follower would land in a position that would adversely affect hand alternation or finger repetition. In this regard, while pressing two keys in sequence, alternation of hands was considered best, followed by rolling out of fingers of the same hand, followed by rolling in of fingers of the same hand and repeating the same finger was considered worst. For rolling out and rolling in, rolling over fewer keys was preferred.

Malayalam uses vowels in both upper case (swara aksharanga) and lower case (swara chihngal), upper case occurring only at the beginning of a word. It would be advantageous if dynamic assignment of vowel symbols to the same position as the vowels themselves could be done by the keyboard's firmware or a software based on the key pressed immediately prior or position of cursor. The most common conjuncts in Malayalam are geminates. If a special geminate key were possible, it would reduce three key-presses needed to type a geminate otherwise, to two. The analysis of frequency done with such a layout in mind was used to design a layout using the same principles given above.

1.4.3 Evaluation of the Layouts

Evaluation of the keyboard layouts was done on the testing portion of the corpus. The test corpus was modified separately for each of the four layouts. The modifications were made to replace that combination of characters which were typed by keys other than the combination of the actual characters, to reflect the actual keys pressed. As an example, geminates were converted to the base character and the special key represented by “:”. Moreover, each press was to correspond to one character. Those keys which correspond to a combination of characters were converted to a Devnagiri letter. For example, “ഃ” of InScript was converted to “ः”. The sed scripts named ‘changestestxxx.sed’ used for those changes are available from the code repository.

The frequency of each character and of consecutive pairs were calculated using standard Linux utilities for each of the modified test corpus separately. R (R Core Team 2021) was used for further analysis after importing this data. Not all Malayalam Unicode characters were present in any given layout. The proportion of Malayalam characters in the corpus not present in a layout was calculated for each of the four layouts. This proportion was used as a correction for some metrics calculated.

The evaluation focused on five questions:

1. What is the difference between each of the four layouts in the number of key-presses (including modifier keys) required to type the same passage? This was answered by calculating the number of key-presses including those of modifier keys required to type the test corpus expressed as a proportion of total Malayalam Unicode characters in the test corpus.

2. Are the two hands used to a similar extent for typing a given passage? In other words, is one of the hands over used? This was answered by expressing the absolute difference between total characters typed by one hand and that by the other hand as a ratio of total characters in the test corpus.
3. Does the work of typing that falls on each of the different fingers correspond with their abilities. In other words, does more work fall on the stronger fingers? This is answered by calculating the sum of frequency of each character typed by each of the different fingers. The number of keypresses for each of the grid rank of the score is an estimate of how well the most preferred positions were used.
4. In typing the characters that each of the fingers have to type, does it move from its resting / home position? We can examine this aspect by dividing the keys into different zones - one for the home keys (home zone), another for positions that fall just one key away from the home position (neighbour zone) and others (town zone) and finding the proportion of characters that fall in each of these zones.

The average rank for each layout, calculated by summing the product of proportion of each rank with the rank will serve as a global measure of how well the layout agrees with the preferred ranks.

5. In typing consecutive characters, do the fingers change in a favourable direction or not? This aspect is analysed by considering overlapping pairs of characters. In a sequence of say “abcde”, the overlapping pairs would be ‘ab’, ‘bc’, ‘cd’ and ‘de’. For each such pair, a SequenceScore of 10 was assigned if the same finger was used to type both characters of the pair and at the same position. If the same finger was repeated but at a different position, more penalty was assigned according to whether the movement was in the preferred direction or not. For all fingers except the little finger, upward movement in the vertical direction i.e., moving up a row was preferred compared to moving down a row. For the little finger penalty was higher for moving up. In the horizontal direction, moving inwards, i.e., moving a column in the direction of thumb while typing was considered preferable. When the same finger was repeated for typing a pair, if the movement was in the preferred direction, an additional sequence penalty equal to the number of rows and columns moved was added. If the direction of movement was opposite to the direction preferred, a penalty of one more than the number of rows and columns moved was added. If different fingers of the same hand were used to type a pair, SequenceScore was assigned as shown in table 1. If modifier keys were used to type one or both of the pair, then the SequenceScore assigned according to the above principles was multiplied by the number of modifier keys used plus one. If different hands were used to type the characters of a pair, a score of 0 was used. Thus, the SequenceScore for a consecutive pair of characters indicates the rank from among the preferred sequence of changes, with lesser preference indicated by a higher number. But the rise in the numerical score assigned is discontinuous with possible gaps of irregular magnitude between adjacent scores. Thus, the SequenceScore has only an ordinal meaning. We summarise the SequenceScores for a layout by the average SequenceScore obtained by summing up the

product of proportion of pairs with a particular SequenceScore with one more than that SequenceScore (one is added to avoid multiplying with zero).

Table 1 SequenceScore assigned for finger change while typing consecutive pairs using the same hand.

first	second				
Thumb	Index	Middle	Ring	Little	
Thumb	10	1	2	3	4
Index	5	10	1	2	3
Middle	6	5	10	1	2
Ring	7	6	5	10	1
Little	8	7	6	5	10

1.5 Results

The corpus used contained 11,70,10,525 characters which reduced to 10,52,42,702 after the initial changes were made. From the 2,06,176 lines in this cleaned corpus, 1,23,705 (60%) were randomly separated into the training corpus and remainder as testing corpus. The training corpus had 6,32,22,411 characters and the testing corpus had 4,20,20,291 characters.

1.5.1 Frequency of Characters

The frequency of characters in the unmodified corpus analysed for the purpose of designing a layout with statically assigned characters is shown in table 2. The characters are sorted in the descending order of frequency. The rank based on the frequency and the proportion of the character in the entire test corpus are shown within parentheses.

Table 2: Frequency of Malayalam characters when analysed unmodified.

്	space	ി	ക	ന	ു	ത	ാ	യ	പ
(1 - 12.97)	(2 - 11.35)	(3 - 6.48)	(4 - 4.95)	(5 - 4.93)	(6 - 4.57)	(7 - 4.46)	(8 - 4.10)	(9 - 3.41)	(10 - 2.62)
ട	ര	വ	മ	െ	ോ	സ	ല	ച	റ
(11 - 2.49)	(12 - 2.31)	(13 - 2.23)	(14 - 2.18)	(15 - 2.08)	(16 - 1.98)	(17 - 1.89)	(18 - 1.81)	(19 - 1.45)	(20 - 1.45)
ള	ണ	ഴ	ബ	േ	ോ	ർ	ൽ	ദ	അ
(21 - 1.34)	(22 - 1.32)	(23 - 1.11)	(24 - 1.04)	(25 - 0.99)	(26 - 0.98)	(27 - 0.96)	(28 - 0.94)	(29 - 0.84)	(30 - 0.68)
ീ	ഗ	ഷ	ശ	ൻ	ൂ	ജ	ശീ	എ	ബ
(31 - 0.60)	(32 - 0.59)	(33 - 0.55)	(34 - 0.55)	(35 - 0.50)	(36 - 0.50)	(37 - 0.46)	(38 - 0.44)	(39 - 0.43)	(40 - 0.42)
ധ	ഹ	ഇ	ഭ	ആ	ണ	ഡ	ഞ	ൊ	ഥ
(41 - 0.40)	(42 - 0.38)	(43 - 0.36)	(44 - 0.34)	(45 - 0.33)	(46 - 0.32)	(47 - 0.30)	(48 - 0.24)	(49 - 0.23)	(50 - 0.23)
ഴ	ഒ	ഉ	ൈ	ഫ	NJ	ൗ	ഖ	ൂ	ഏ
(51 - 0.22)	(52 - 0.22)	(53 - 0.21)	(54 - 0.18)	(55 - 0.18)	(56 - 0.15)	(57 - 0.13)	(58 - 0.12)	(59 - 0.11)	(60 - 0.08)
ൺ	ഓ	ഘ	ൌ	ഠ	ഐ	ഛ	ഞ	ഊ	ഃ
(61 - 0.08)	(62 - 0.06)	(63 - 0.06)	(64 - 0.05)	(65 - 0.04)	(66 - 0.02)	(67 - 0.01)	(68 - 0.01)	(69 - 0.01)	(70 - 0.00)
ഡ	ഋ	ൠ	ൡ	ം	ർ	ൣ	വ	൶	൷
(71 - 0.00)	(72 - 0.00)	(73 - 0.00)	(74 - 0.00)	(75 - 0.00)	(76 - 0.00)	(77 - 0.00)	(78 - 0.00)	(79 - 0.00)	(80 - 0.00)
ൺ	൹	ൺ	ൻ	ർ	ൽ	ൾ	ൿ	ൿ	ൿ
(81 - 0.00)	(82 - 0.00)	(83 - 0.00)	(84 - 0.00)	(85 - 0.00)	(86 - 0.00)	(87 - 0.00)	(87 - 0.00)	(88 - 0.00)	(89 - 0.00)
ൿ	ൿ	ൿ	ൿ						
(90 - 0.00)	(90 - 0.00)	(91 - 0.00)	(91 - 0.00)						

Shaded cells indicate non-printing characters.

NJ = zero width non joiner

Analysis of the frequency of overlapping consecutive pairs of characters in the unmodified test corpus is shown in table 3. Not all pairs are shown. Pairs in which one of the character was space were excluded. After sorting the pairs in descending frequency, pairs were scrutinised in sequence for inclusion in table 2. A pair was selected only if at least one of its constituent was not selected earlier. Thus, for any character its most frequent pair is shown. However, not all

characters are shown in table 2, though all of the 69 characters in regular use except ൠ are included. Rank and proportion of each pair out of all pairs are shown in parentheses.

Table 3: Frequency of important leader follower pairs of Malayalam characters when analysed unmodified.

ന - ൠ (1 - 1.84)	ത - ൠ (2 - 1.83)	ക - ൠ (4 - 1.68)	ത - ി (12 - 1.03)	പ - ൠ (14 - 0.91)	ഃ - ട (15 - 0.85)	ഃ - റ (16 - 0.83)	ഃ - ള (18 - 0.82)
ണ - ൠ (20 - 0.79)	ു - ൠ (22 - 0.78)	സ - ൠ (24 - 0.76)	യ - ി (25 - 0.75)	ാ - യ (31 - 0.68)	മ - ാ (34 - 0.65)	ി - ല (35 - 0.65)	ബ - ൠ (40 - 0.58)
ി - ി (42 - 0.57)	ര - ൠ (43 - 0.57)	ച - ൠ (44 - 0.54)	വ - ി (49 - 0.49)	ദ - ൠ (68 - 0.35)	ട - ി (69 - 0.34)	ഃ - ള (70 - 0.33)	ബ - ൠ (76 - 0.32)
എ - ന (86 - 0.25)	ക - ി (126 - 0.17)	ഃ - ഡ (127 - 0.17)	ദ - ി (129 - 0.16)	ഃ - ഷ (135 - 0.16)	ദ - റ (137 - 0.16)	ാ - ി (138 - 0.16)	ഗ - ൠ (140 - 0.15)
ത - ൠ (150 - 0.14)	ഃ - ഡ (159 - 0.14)	ഃ - ഡ (167 - 0.13)	യ - ി (168 - 0.13)	ാ - ി (169 - 0.13)	ശ - ൠ (170 - 0.13)	ഭ - ാ (176 - 0.11)	അ - വ (181 - 0.11)
ു - ട (183 - 0.11)	ക - ി (189 - 0.10)	ഇ - ത (206 - 0.09)	ബ - ൠ (211 - 0.09)	ഡ - ൠ (216 - 0.08)	ി - ക (219 - 0.08)	ജ - ൠ (232 - 0.08)	ഴ - ി (244 - 0.07)
േ - ഹ (264 - 0.07)	ഫ - ൠ (273 - 0.07)	ഉ - പ (312 - 0.06)	ു - ത (322 - 0.06)	ആ - ദ (398 - 0.04)	വ - ൠ (459 - 0.03)	ഘ - ട (473 - 0.03)	ഏ - ള (509 - 0.03)
ോ - ണ (534 - 0.03)	ൈ - ന (545 - 0.02)	പ - റ (656 - 0.02)	ഓ - ഫ (690 - 0.02)	സ - ി (820 - 0.01)	ഃ - ഡ (1087 - 0.01)	ഐ - ക (1119 - 0.01)	ഔ - ദ (1238 - 0.00)
ഊ - ി (1364 - 0.00)	ഈ - സ (1404 - 0.00)	ന - ി (1503 - 0.00)	ു - ഡ (1684 - 0.00)	ഋ - ഷ (1847 - 0.00)	ഃ - ി (1847 - 0.00)	യ - ാ (1945 - 0.00)	ി - ി (1977 - 0.00)

The frequency of characters in the modified corpus analysed for the purpose of designing a layout with dynamically assigned characters is shown in table 4. The characters are sorted in the descending order of frequency. The rank based on the frequency and the proportion of the character in the entire test corpus are shown within parentheses.

Table 4: Frequency of Malayalam characters when analysed after modifications.

space (1 - 11.90)	ഇ (2 - 7.17)	അ (3 - 7.11)	gem (4 - 6.45)	ഉ (5 - 5.02)	ആ (6 - 4.65)	ന (7 - 4.35)	ക (8 - 4.01)	ത (9 - 3.56)	യ (10 - 3.50)
chillu (11 - 3.06)	ല (12 - 2.71)	എ (13 - 2.63)	റ (14 - 2.52)	ര (15 - 2.43)	വ (16 - 2.32)	പ (17 - 2.27)	മ (18 - 2.20)	ട (19 - 2.17)	ഠ (20 - 2.08)
സ (21 - 1.94)	ള (22 - 1.59)	ണ (23 - 1.42)	ഏ (24 - 1.12)	ഓ (25 - 1.09)	ച (26 - 1.05)	ഭ (27 - 0.79)	ഈ (28 - 0.76)	ബ (29 - 0.61)	ഗ (30 - 0.60)
ഷ (31 - 0.57)	ശ (32 - 0.56)	ഊ (33 - 0.53)	ഓ (34 - 0.47)	ജ (35 - 0.47)	ബ (36 - 0.44)	റ്റ (37 - 0.42)	ധ (38 - 0.42)	ഹ (39 - 0.40)	ഭ (40 - 0.36)
ന്റ (41 - 0.33)	ഡ (42 - 0.31)	ഥ (43 - 0.24)	ഴ (44 - 0.23)	ഐ (45 - 0.22)	ഘ (46 - 0.19)	ഞ (47 - 0.17)	NJ (48 - 0.15)	ഖ (49 - 0.12)	ഘ (50 - 0.12)
ഔ (51 - 0.07)	ഘ (52 - 0.06)	ഓ (53 - 0.04)	ഔ (54 - 0.01)	ഔ (55 - 0.00)	ഔ (56 - 0.00)	ഔ (57 - 0.00)	ഔ (58 - 0.00)	ഔ (59 - 0.00)	ഔ (60 - 0.00)
ഔ (61 - 0.00)	ഔ (62 - 0.00)	ഔ (63 - 0.00)	ഔ (64 - 0.00)	ഔ (65 - 0.00)	ഔ (66 - 0.00)	ഔ (67 - 0.00)	ഔ (68 - 0.00)	ഔ (69 - 0.00)	ഔ (70 - 0.00)
ഔ (71 - 0.00)	ഔ (71 - 0.00)	ഔ (72 - 0.00)	ഔ (73 - 0.00)	ഔ (74 - 0.00)	ഔ (75 - 0.00)	ഔ (75 - 0.00)	ഔ (76 - 0.00)	ഔ (76 - 0.00)	

Shaded cells indicate nonprinting characters.

gem = key to geminate a character

chillu = key to remove inherent vowel of a consonant

NJ = zero width non joiner

Analysis of the frequency of overlapping consecutive pairs of characters in the modified test corpus is shown in table 5. Not all pairs are shown. After sorting the pairs in descending frequency, pairs were scrutinised in sequence for inclusion in table 5. A pair was selected only if at least one of its constituents was not selected earlier. Thus, for any character its most frequent pair is shown. However, not all characters are shown in table 5, though all of the 69 characters in regular use are included. Rank and proportion of each pair out of all pairs are shown in parentheses.

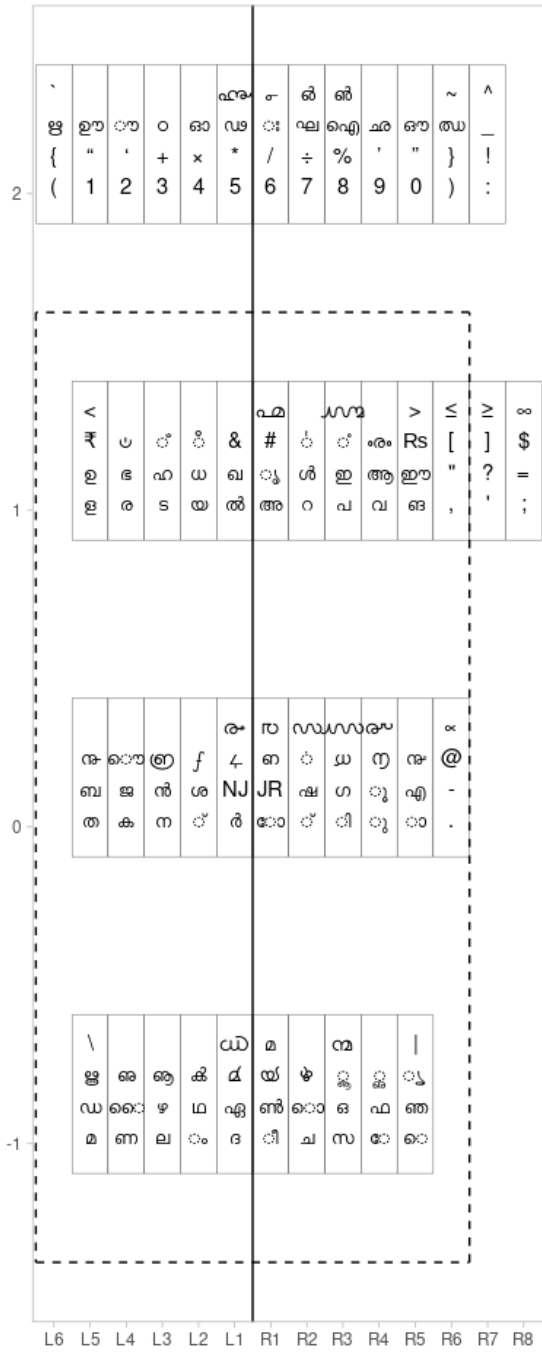
Table 5: Frequency of important overlapping consecutive pairs of Malayalam characters when analysed after modification.

gem - ഉ (5 - 1.36)	ന - gem (6 - 1.35)	ഇ - ല (7 - 1.30)	ക - gem (9 - 1.18)	ത - gem (11 - 1.11)	റ - chillu (13 - 1.00)	അ - റ (16 - 0.92)	ഉ - ൠ (19 - 0.82)
യ - ഇ (22 - 0.79)	ണ - അ (23 - 0.77)	സ - അ (25 - 0.75)	ആ - യ (26 - 0.75)	മ - ആ (33 - 0.67)	ര - ഉ (37 - 0.59)	വ - ഇ (40 - 0.52)	അ - ട (46 - 0.49)
പ - gem (47 - 0.48)	ബ - gem (48 - 0.48)	ച - gem (49 - 0.48)	ള - chillu (53 - 0.46)	gem - എ (64 - 0.38)	ദ - അ (81 - 0.28)	ന്റ - എ (85 - 0.27)	ഒ - റ (118 - 0.19)
അ - ഫ (123 - 0.18)	അ - ഷ (125 - 0.17)	അ - ധ (127 - 0.16)	ഗ - അ (146 - 0.15)	gem - ഏ (147 - 0.15)	അ - NJ (160 - 0.13)	യ - ഓ (161 - 0.13)	ശ - അ (167 - 0.12)
ഭ - ആ (168 - 0.12)	ഊ - ട (176 - 0.11)	അ - റ്റ (183 - 0.11)	ഈ - ക (223 - 0.09)	ഡ - അ (225 - 0.09)	ബ - അ (228 - 0.09)	ജ - ന (244 - 0.08)	ഞ - gem (245 - 0.08)
ഴ - ഇ (251 - 0.08)	ഏ - ഹ (275 - 0.07)	ഫ - അ (286 - 0.07)	ള - ത (322 - 0.06)	ഖ - അ (457 - 0.04)	ഐ - ന (467 - 0.03)	ഘ - ട (478 - 0.03)	പ - റ (637 - 0.02)
ഔ - ണ (750 - 0.01)	അ - ഛ (990 - 0.01)	ഊ - ഡ (1438 - 0.00)	അ - ൠ (1558 - 0.00)	യ - ആ (1624 - 0.00)	ഇ - ക്ക (1650 - 0.00)	ആ - റ്റ (1788 - 0.00)	അ - അ (1790 - 0.00)

1.5.2 New Keyboard Layout Designs

The new layouts designed using the data from frequency analysis are shown in figure 3. The layout designed for static assignment of characters is named Puthunira Aarambham (പുതുനീര അരംഭം) and the one designed for dynamic assignment of characters is named Puthunira Charutha (പുതുനീര ചാരൂത).

Puthunira Aarambham



Puthunira Charutha

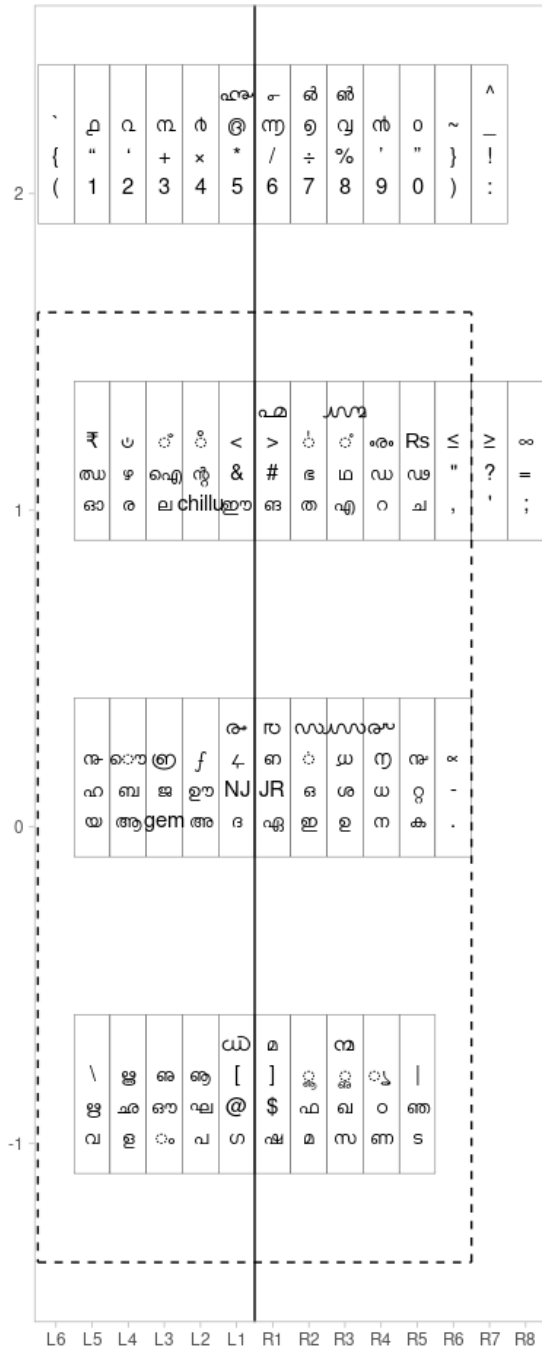


Figure 3: Puthunira Aarambham and Puthunira Charutha layouts

1.5.3 Evaluation of Layouts

1.5.3.1 Average key press per Unicode character

Figure 4 shows the result of comparison of the four layouts tested with regards to the average number of key-presses including modifier keys per Unicode character in the test corpus, corrected for the characters not represented in a layout.

=====

Comparison of layouts by average keypress per unicode character



Figure 4: Comparison of layouts by average key-press per Unicode character.

Result of comparison of layouts based on imbalance between the use of the two hands for the primary key-press is shown in figure 5.

Comparison of layouts by left right imbalance of primary key press

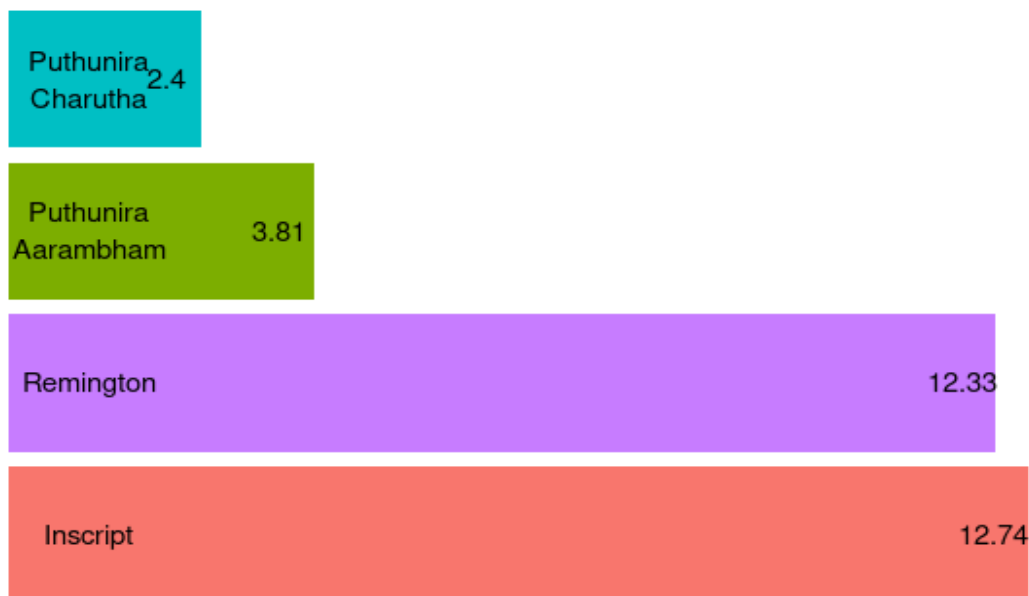


Figure 5: Comparison of layouts by left right imbalance of primary key press.

1.5.3.2 Finger Usage

The number of characters that fall on each finger at each combination of row and column is shown in figure 6. The numbers include characters in all levels - with and without modifiers. The numbers are shown in lakhs. The rows are distinguished by colours and named in the legend. The home row is the row where the finger rests when not typing. The bottom row is one row below, the top row is one row above, and number row is two rows above.

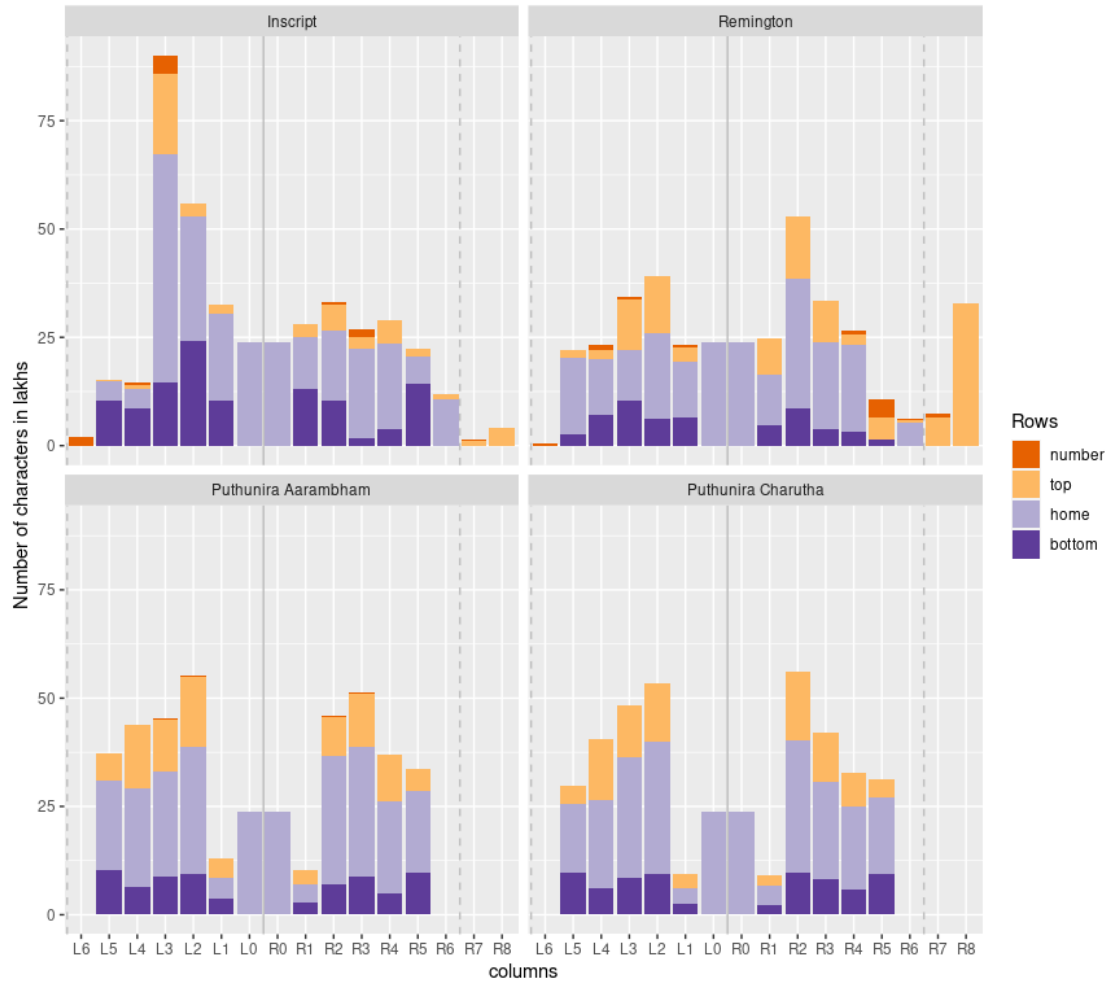


Figure 6: Finger usage pattern of the layouts.

1.5.3.3 Finger Positions

The proportion of key press that falls on each of the three zones is shown in table 6. The average rank for each of the layouts is also shown in table 6. The best numbers in proportions that fall in home zone and for the average rank are highlighted.

Table 6: The proportion of key-presses that falls on the different zones of the layout.

Layouts	Percentage of characters in each zone			Average Rank
	Home	Neighbour	Town	
InScript	48.36	48.09	3.55	11.20
Remington	46.62	40.70	12.67	15.83
Puthunira Aarambham	58.16	41.62	0.22	7.57
Puthunira Charutha	57.87	42.13	0.00	6.81

1.5.3.4 Finger Change

The distribution of SequenceScore for each of the four layouts is shown in figure 7. In the graph, the proportion of SequenceScore of value zero is not shown as it was the most frequent SequenceScore (by a very high margin) for all four layouts and including it would make interpreting the distribution of other SequenceScores difficult.

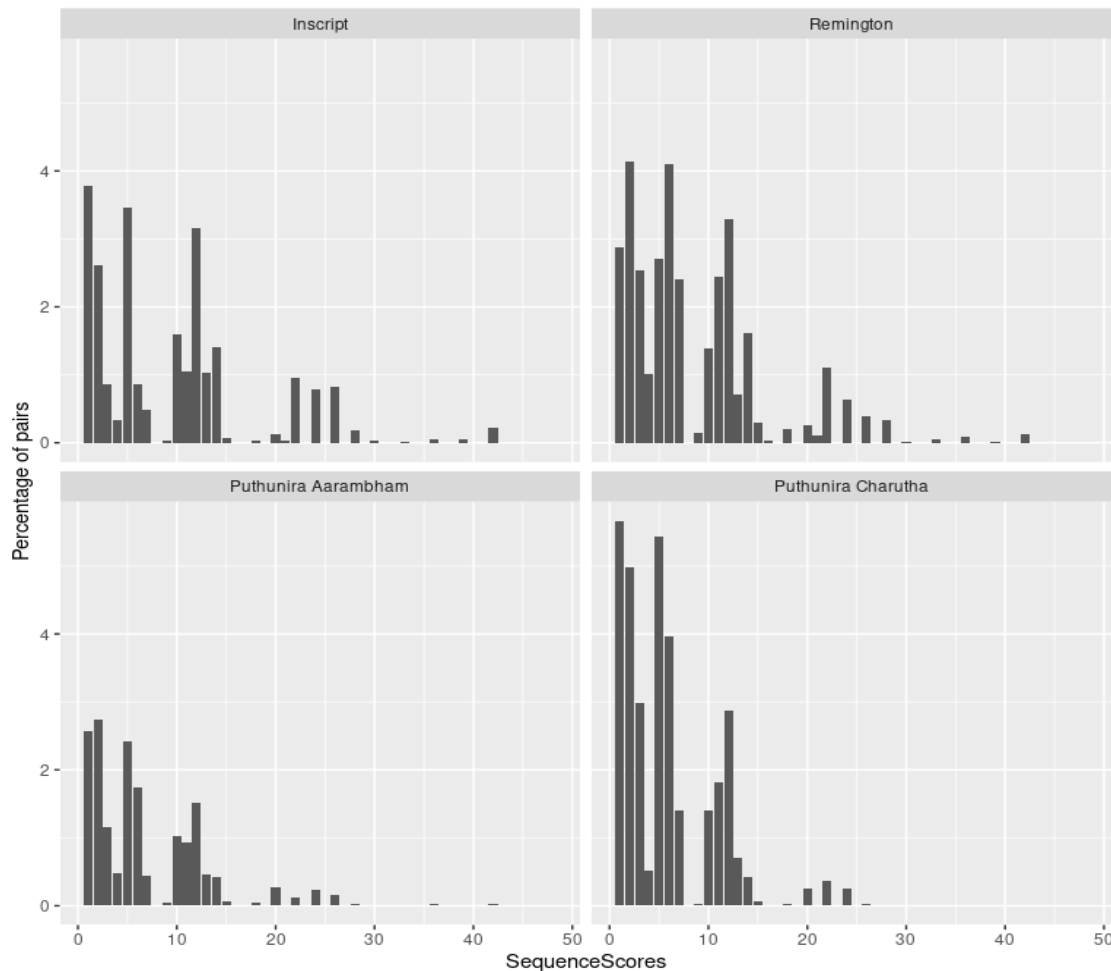


Figure 7: Finger change pattern of the layouts.

Table 7 shows the worst SequenceScore each layout has, the percentage of SequenceScore of zero out of all consecutive pairs and the average SequenceScore. The percentage of pairs for which a SequenceScore could not be calculated (because one or more of the pair was not represented in the layout) are also shown.

Table 7: Layouts compared by SequenceScore

Layouts	Percentage not included in the layout	Worst Sequence-Score	Percentage of a Sequence-Score of zero	Average Sequence-Score
Inscript	0.003	45	76	3.24
Remington	0.002	48	67	3.77
Puthunira Aarambham	0.001	45	83	2.13
Puthunira Charutha	0.002	42	67	2.94

1.6 Discussion

The training corpus was modified prior to character frequency analysis. Two types of changes were made. First set of changes were aimed for “cleaning” the corpus by eliminating characters not relevant for typing Malayalam, removing extra white spaces, and bringing uniformity in the way chillu and the conjunct റ്റ were represented in the corpus. The second set of changes were intended to rectify the bias in frequencies that are caused by the restrictions imposed by fonts in typing the conjuncts involving റ / റ and ല / ല. The underlying assumption for these “corrections” is that the keyboard layouts developed from these analyses would cater better to circumstances that permit typing all conjuncts permitted by the language.

The ranks assigned to each of the key positions in the idealised grid of keys is in no way objective. It is a subjective ranking of the author. However, the principles followed in ranking should be agreeable to most - The stronger fingers (thumb, index, middle) were preferred over weaker (ring, little). Home positions were preferred over other positions which required finger movements. Positions that required less finger travel were preferred compared to those which required more finger travel from home positions.

Similarly, the SequenceScore used to assess the change of fingers is also a subjective score. However, it would be agreeable to most that repeating a finger would be the least preferred as the finger doesn’t rest between typing a consecutive pair and that alternating the hands would be the most ideal.

The layout Puthunira Aarambham makes no assumption other than that a key press will result in a Unicode character. The most frequent character is chandrakala ്, nearly 13% of

characters, about double the frequency of “ റ്റ ” the Malayalam character second highest in frequency. As the gulf between the first and second commonest Malayalam characters was too wide, chandrakala is assigned to both hands, to the index fingers. This is a conscious design decision to reduce imbalance between the hands and to improve SequenceScore.

The layout Puthunira Charutha makes many assumptions. The first assumption is that the keys for the vowels will be used as such for the vowel signs, the assignment being made dynamically by the computer. This will save 13 key positions. Also, there would be lesser key positions to remember. The next assumption made is that a position will be marked for a special key to cause gemination of base consonants. This will result in two key presses for a geminate (the base character plus the special geminate key) instead of three (base character followed by chandrakala followed again by the base character). Considering that the geminates are frequent in Malayalam, this could reduce the key-presses substantially. The reason for including a special key position to change a consonant into its vowel-less form (chillu) is not based on ease of typing. In fact, it may slightly increase the keystrokes required. The reason for a “chillu key” is to shield the users from the need to decide from among the two ways in which chillu can be encoded in Unicode. Moreover, it can help reduce the confusion brought by Unicode representation of chillus as characters different from the base characters. Being able to produce a chillu from a base character by pressing “chillu key” would help reinforce the fact that chillu is just a vowel-less form of the consonant rather than a new character. The price to pay for “chillu key” should be an extra key-press for all the chillus. However, from the frequency analysis it can be seen that only റ and റ്റ would have fallen in the first level (without modifier keys) and all other chillu even otherwise would have needed two key-presses. Thus, the cost of a special “chillu key” is an extra key-press for റ and റ്റ. The Unicode representation of the conjunct റ്റ has been revised many times. It has been assigned to a sequence derived from its morphology rather than the conceptual constituents. Assigning the റ്റ to a separate key can help shield the users from the multiple representations of റ്റ. If the representation changes in future too, the users will be shielded from such changes. Similarly, it was thought better to assign a separate key for റ as its current Unicode representation is based on morphology.

For both the layouts designed, only the first and second level of characters were assigned based on frequency analysis. The third level was used for the archaic and rarely used characters. The fourth level was used for the Malayalam fractions. Malayalam digits are present only in the Puthunira Charutha. They are presumed to be assigned to the second or third levels of numerical keypad of Puthunira Aarambham. Both layouts assign keys for zero width joiner and zero width non-joiner. The punctuation marks and symbols are placed in the number row and outside the 5+5 columns in the other rows. This is decidedly different from the prevailing assignment in QWERTY layout.

The expectations and assumptions of the new layouts were evaluated against the two commonly used Malayalam layouts InScript and Remington. Before we consider the evaluation results, we need to take a look at InScript and Remington layouts. One design principle of InScript (“A Document for Enhanced InScript Keyboard Layout 5.2” (2010)) is that it is similar across the different Indian languages. It should be clear that the frequency of any one character in any given language would not match the usage frequency of the same character in another language. This design principle of InScript makes it challenging for that layout to provide typing ease in any language (barring the first language for which it was designed). Remington layout was originally designed for typewriters (“Remington” (2010)). Though frequency of glyphs were taken into account, the scarcity of key positions was a real problem for its design. To circumvent the problem, some characters were sacrificed, introducing some ad hoc methods to overcome their absence. When the Remington layout was adopted for computers, the missing characters were added to whatever positions were free, without taking into account their frequency. Remington layout assigns some geminates to individual keys, thus making it possible that the keystrokes required would be less.

The first evaluation metric that we consider is the number of keystrokes that would be required to type the test corpus. It is expressed as the average number of strokes required to type one Unicode character. The keystroke count includes the modifier keys. Though the number of characters not represented in a layout is very small for all four layouts, it is still used to correct the evaluation metric. The layout with the lowest number of keystrokes per Unicode character is Puthunira Charutha which requires only 1.014 keystrokes on an average. The InScript layout has the worst score of 1.126. Though Puthunira Aarambham does not assign multiple characters to a single key, it has a better score than InScript because it assigns more frequently assigned characters to the first level which doesn’t use modifier keys. The Puthunira Charutha requires 10% less keystrokes compared to InScript.

The second metric to consider is the hand imbalance. It is expressed as the proportion of characters that any one hand has to type more than the other. This metric too was corrected for characters not represented in a layout. The Puthunira Charutha has the least imbalance of only 2.4%. The Puthunira Aarambham follows close behind Puthunira Charutha and much ahead of both InScript and Remington. InScript has the highest imbalance 12.7% - to mean that the left hand does about 12.7% more typing than the right.

The third aspect to consider is how the work of typing is shared by the different fingers. This is shown in the stacked bar diagram of figure 7. For each column of keys, the number of characters that fall on that column is shown. The numbers for each row in each column is colour coded and stacked. The graph for InScript shows that a disproportionate number of characters fall on L3 column, left middle finger; almost triple the number that falls on R2 or R3. The index finger column L1 and R1 types nearly as much as L2 and R2. This is not desirable as the L1 and R1 columns require moving the finger inwards. Also, the number typed by the left middle finger

by stretching to the topmost row is significant, clearly visible in the graph. The Remington layout assigns a disproportionate number of characters to the right little finger and that too at a position where it has to stretch out and upwards as seen in the bars for R8 and R7.

Clearly, the distribution of typing effort to the various fingers is more balanced in both Puthunira Aarambham and Puthunira Charutha as seen in the graphs. The graphs for both layouts show decreasing height of bars from the home column of index key to that of the little finger on both sides. The positions where fingers are to be moved are used much less than home positions. No characters are typed by stretching the right little finger to the most extreme columns. The use of the number row is limited in Puthunira Aarambham and absent in Puthunira Charutha.

While the graphs give a comprehensive view of typing effort across fingers, some numerical summaries given in table can make the picture clearer. When we divide the key positions into zones, we see that both Puthunira Aarambham and Puthunira Charutha uses the home positions nearly 60% of time while for InScript and Remington, home positions are used only under 50% of time. The zone more than one key away from home positions (which I call town zone), is not used by Puthunira Charutha and is used only to a very small extent by both Puthunira Aarambham and InScript. However, Remington uses this zone more than 12% of the time.

Assuming that the ranks assigned to key positions is acceptable, the average rank can be considered as a more comprehensive metric indicating the fit of the matter typed to the preference expressed as ranks. The Puthunira Charutha has the best overall (lowest) rank, closely followed by Puthunira Aarambham.

The final aspect to consider is how the fingers change when typing consecutive characters. The best would be to alternate between hands so that there is a period of rest for the fingers. Hence, subjective scores are given to the various combination of movements with the lowest number for alternation of hands and the highest number for repetition. Penalty is increased if modifier keys are used for one or both characters of a pair. Figure 6 shows the proportion of consecutive pairs for each of the SequenceScores. A score of zero is the commonest, accounting for 83% of pairs in Puthunira Aarambham. The lowest proportion for a SequenceScore of zero is for Puthunira Charutha; but even there it constitutes more than 66%. This proportion requires cautious interpretation relating it to the total number of keystrokes. Thus, though the proportion of pairs where hands are alternated is lower for Puthunira Charutha, the total number of pairs that need to be typed is lower for Puthunira Charutha.

The graphs of figure 8 exclude the proportion for a sequence score of zero. From the distribution of the remaining SequenceScores, it can be seen that for Puthunira Charutha the better (lower numeral) scores have higher bars. Similar is the case with Remington, though not as high as the bars of Puthunira Charutha. All layouts except Puthunira Charutha have

SequenceScore beyond 40. Thus, though the proportion of zero SequenceScore is lower for Puthunira Charutha, the layout is easier on those pairs that does not involve alternation of hands.

The numerical summaries presented in table 7 bring greater clarity. The proportion of pairs not type-able by a layout is low for all layouts, least for Puthunira Aarambham. The worst SequenceScore for a layout is the lowest for Puthunira Charutha. The average SequenceScore is also shown. This metric can be considered as an overall measure of how well the distribution of finger changes matches with the preferred distribution of such changes. Puthunira Aarambham has the best score in this measure, followed by Puthunira Charutha.

1.7 Conclusion

A keyboard layout designed taking into account the frequency of character usage has the potential to make typing easy. The lack of such a layout for Malayalam is tackled in this study to create two layouts named Puthunira Aarambham and Puthunira Charutha. The Puthunira Aarambham is a simple layout that aims to lay out the characters in a manner efficient for typing. The Puthunira Charutha aims to reduce typing effort even more by envisaging dynamic assignment of vowels and their symbols. It also aims to reduce some of the effects of some restrictions introduced by the Unicode encoding of Malayalam on the language.

Objective evaluation of the new layouts confirms that the new layouts does reduce typing effort in comparison with two of the prevalent layouts InScript and Remington. They are better in needing lesser number of key presses, better balance in using the two hands, better distribution of typing effort across the different fingers and across different positions of the same fingers and in having a favourable distribution of change of fingers while typing consecutive characters. While Remington requires lesser keystrokes than Puthunira Aarambham, it has a far worse distribution of typing effort across the different fingers and positions. InScript requires the most key presses and has an unimpressive distribution of typing effort across the different fingers and positions, though less severe than Remington. Puthunira Aarambham is better than Puthunira charutha in only the finger change pattern.

The best layout considering all the factors is Puthunira Charutha. However, its implementation may be difficult. When implemented, it has the potential to be learned faster as it requires a smaller number of key positions to remember. If it cannot be implemented, Puthunira Aarambham is the next option. InScript is to be favoured only by those who would type more than one Indian language. In this situation, the common layout across languages would be an advantage, though it would sacrifice ease of typing.

1.8 End Notes

The computer code used in this study, the complete results of frequency analysis and the implementation code for Puthunira Aarambham and Puthunira Charutha are available at the git repository of the author available at <https://gitlab.com/ajithramayyan/puthunira>.

1.9 Acknowledgments

I would like to record my sincere thanks to Padmini R, who assisted me with designing the layouts using the frequency analysis result. The idea of dynamic assignment of vowel symbols and vowels to the same keys was inspired by Google’s Gboard Android keyboard. The idea of a separate key to mark gemination is expressed in Nair (1971). Though the idea of a special geminate key was not inspired by Nair (1971), and the function of the geminate key of Nair (1971) is to only apply a mark to show gemination, I acknowledge here that had there been the technical advantage of computers at the time of study of Nair (1971), it would have led to a geminate key as envisaged in this study. I want to record my sincere gratitude to Tim of Cross Validated for showing me how best to summarise rank preferences.

=====

Bibliography

- “A Document for Enhanced Inscript Keyboard Layout 5.2.” 2010. CDAC.
Bucaco, OJ. 2010. “The Workman Keyboard Layout Philosophy.” Available at <https://viralintrospection.wordpress.com/2010/09/06/a-different-philosophy-in-designing-keyboard-layouts/> (2022/12/01).
- Capewell, Michael. n.d. “Alternative Keyboard Layouts.” Available at http://www.michaelcapewell.com/projects/keyboard/layout_capewell.htm (2022/12/01).
- Gillespie, Patrick. n.d. “Keyboard Layouts 101.” Available at <http://patorjk.com/keyboard-layout-analyzer/#/about> (2022/12/01).
- Government of Kerala. 1971. “Malayalam Script - Adoption of New Script for Use-Orders Issued.”
- Gutman, Alejandro, and Beatriz Avanzati. 2013. “Malayalam.” Available at <https://languagesgulper.com/eng/Malayalam.html> (2022/12/01).
- “INDIAN SCRIPT CODE FOR INFORMATION INTERCHANGE - ISCII.” 1991. BUREAU OF INDIAN STANDARDS.
Kerala Government Information Technology Department. 2008. “GO(MS) 31/08/ITD.”
- Krzywinski, Martin. n.d. “Evaluating Keyboard Layouts.” Available at http://mkweb.bcgsc.ca/carpalx/?keyboard_evaluation (2022/12/01).
- “Malayalam Corpus by Swathanthra Malayalam Computing.” 2019. Available at <https://github.com/smc/corpus> (2022/03/13). 2019.

=====

Language in India www.languageinindia.com ISSN 1930-2940 23:7 July 2023

Ajith R. MBBS, DCP

PuthuNira – A Set of New Keyboard Layouts for Malayalam

Nair, Bhaskaran. 1971. “പുതിയ മലയാളം അച്ചുഴുത്തുപദ്ധതി.” മാതൃഭൂമി.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.


Ralminov. 2008. “എങ്ങനെ പുതിയ കീബോഡ്‌ലേൗട്ട്‌കൊണ്ട് (വിന്ഡോസ്)?” Available at https://ralminov.wordpress.com/2008/09/10/how_to_create_custom_keyboard_layout/ (2022/12/01).

Ramakumar V. 2004. *വി രാമകുമാറിന്റെ സമ്പൂർണ്ണ മലയാള വ്യാകരണം*. 2nd ed. Siso Books.

“Remington.” 2010. Available at <https://wiki.smc.org.in/Remington> (2022/12/01).

“The Unicode® Standard Version 15.0 – Core Specification.” 2022. The Unicode Consortium.

“ThoolikaUnicode.” 2006. SuperSoft Computer Software Research & Development Centre; Available at <https://www.supersoftweb.com/ThoolikaUnicode.aspx> (2022/12/01).

	<p>Ajith R. MBBS, DCP Nirvriti, T.C. 5/3457(3), Kezhuvankara Pongumoodu, Medical College P.O. Thiruvananthapuram, Kerala 695 011 ajith@disroot.org +91 89390 01898</p>
--	--