Optimization of Thamil Phonetic Keyboard

Sendhil kumar Cheran, Thuraiappah Vaseeharan, Elango Cheran

Abstract

The current standard for the Thamil phonetic keyboard layout is the Tamilnet99 keyboard, which was recommended by a special committee at the 1999 Tamilnet Conference (உலகத் தமிழ் இணைய கருத்தரங்கு மாநாடு) held in Chennai and later accepted by the Government of Thamilnadu [1-2]. An important goal in deciding the optimum phonetic keyboard layout includes placement of Thamil letters with a higher frequency of usage at 'stronger' key positions, and placement of Thamil letters with a lower frequency of usage at 'weaker' key positions [3]. There is reason to believe that improvements can be made in the Tamilnet99 keyboard.

We hypothesize that slight modifications to the Tamilnet99 keyboard layout will increase optimization of the aforementioned goal. The aim of this paper is to analyze whether and how the efficiency and user-friendliness of key positions in the Tamilnet99 keyboard can be improved to match strength of key position with frequency of Thamil letter usage.

Introduction

Two of the most widely-used keyboard styles for inputting Thamil text on the Internet are the Thamil typewriter keyboard (known to many as the Bamini keyboard) and the English transliteration keyboard (typified by Murasu Anjal).

In the not-too-distant future, another major keyboard called the Phonetic keyboard will become the standard keyboard for future generations to type Thamil text. The phonetic keyboard improves on previous keyboards by reducing the number of keystrokes required to input Thamil text, making this keyboard the fastest one. The phonetic keyboard is also the easiest one to learn for people without prior knowledge of English typing [1-3].

Well-known efforts at optimization of keyboard letter placement have been undertaken for the English keyboard, most notably by Dvorak [4-5]. However, to our knowledge there is a paucity of published research on the relative usage frequency of the 247 letters in the Thamil language. Without the availability of these data, it is unclear whether the Tamilnet99 keyboard meets objective criteria for optimal phonetic keyboard layout.

Methods

We created a software program in Python to count Thamil letter occurrences in several representative Thamil texts from the author Kalki and the Project Madurai literary database [6]. The software counts the frequency of occurrences of each of the 247 letters of the alphabet from a UTF-8 encoded text input file. The program outputs to a .html file in the form of a 13 x 18 alphabet chart, with 18 rows of consonants and 13 columns of ìvowelsî (12 columns of the traditional vowels, with the last column being the ஆய்த எழுத்து).

The number of times each consonant is used in the text was calculated by summing the rows of the chart. Each individual row total indicates the number of times the consonant

appears in the text, either as a மெய் or உயிர்மெய் letter. Similarly, column totals were summed to calculate the number of times each vowel was used in the text, either as a உயிர் or உயிர்மெய் letter (and in the case of the ஆய்த எழுத்து-, the number of times it was used alone or as a மெய். Examples of "உயிர்" occurrences of the ஆய்த எழுத்து- are பஃறுளி and எஃகு, and க் and வ் are examples of a மெய்). The row totals were used to order the frequency of consonant usage in the text, and column totals were used to order the frequency of vowel usage.

Next, the total number of keystrokes for each text was calculated for two different keyboards, the Tamilnet99 keyboard and an 'alternate keyboard'. The 'alternate keyboard' is a slight modification of the Tamilnet99 keyboard that replaces all of the consonants with மெய் consonants. Thus, whereas the Tamilnet99 keyboard has அகர உயிர்மெய் letters for the consonants, the 'alternate keyboard' consists of only மெய் letters for consonants on the right hand side of the keyboard. Of note, the 'alternate keyboard' does not use any special Keystroke sequences.

Results

Initial data analysis was performed on subset of modern prose consisting of specific works from the first 100 works of the Project Madurai database along with 4 novels by the novelist ëKalkií Krishnamurthy. The following finished works were selected for the initial data analysis: MP58, MP65, MP66, MP82, MP88, MP97, MP98, MP99, சிவகாமியின் சபதம் (SS), அலையின் ஓசை (AO), பார்த்திபன் கனவு (PK), and பொன்னியின் செல்வன் (PS).

For the entire set of works, the frequency of Single consonant keystrokes is equal to the sum of either மெய் or உயிர்மெய் occurrences of consonants (**Figure 1, Table 1**).

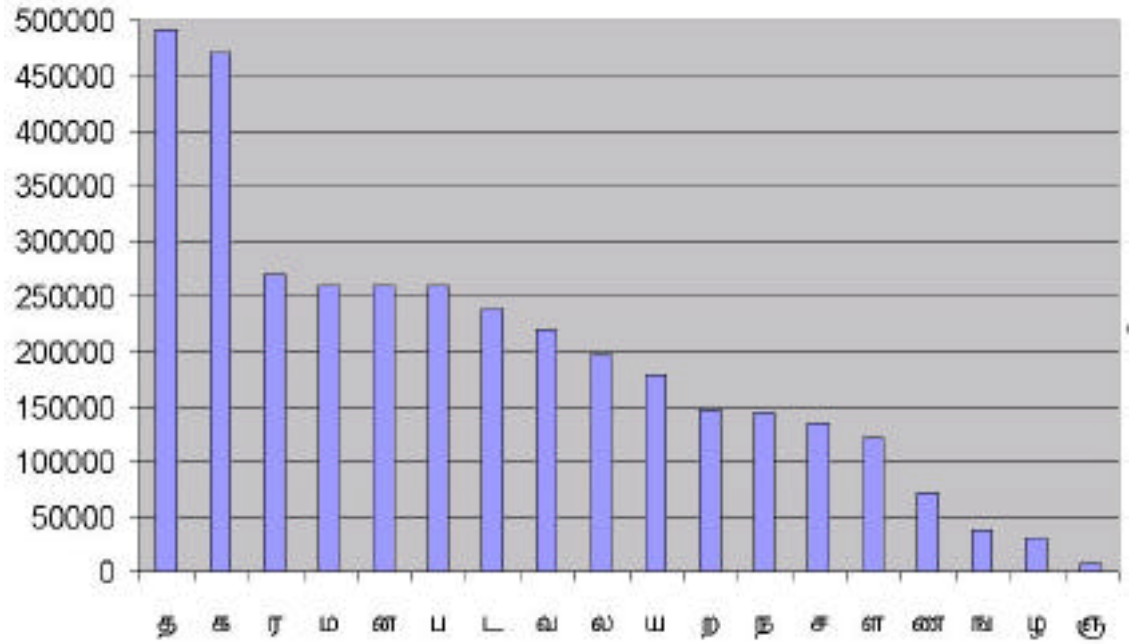

Figure 1. Frequency of Consonant Keystrokes

The frequency of Vowel Occurrences is equal to the sum of உயிர் OR உயிர்மெய் occurrences for the vowels (அ, ஆ, இ, ... , ஓ, ஔ) PLUS the sum of "உயிர்" OR மெய் occurrences for the

ஆய்த எழுத்து (**Figure 2, Table 2**). Except for the letters அ and ஃ, the other vowel occurrences equal the number of vowel keystrokes for both the Tamilnet99 keyboard and ëalternateí keyboard.

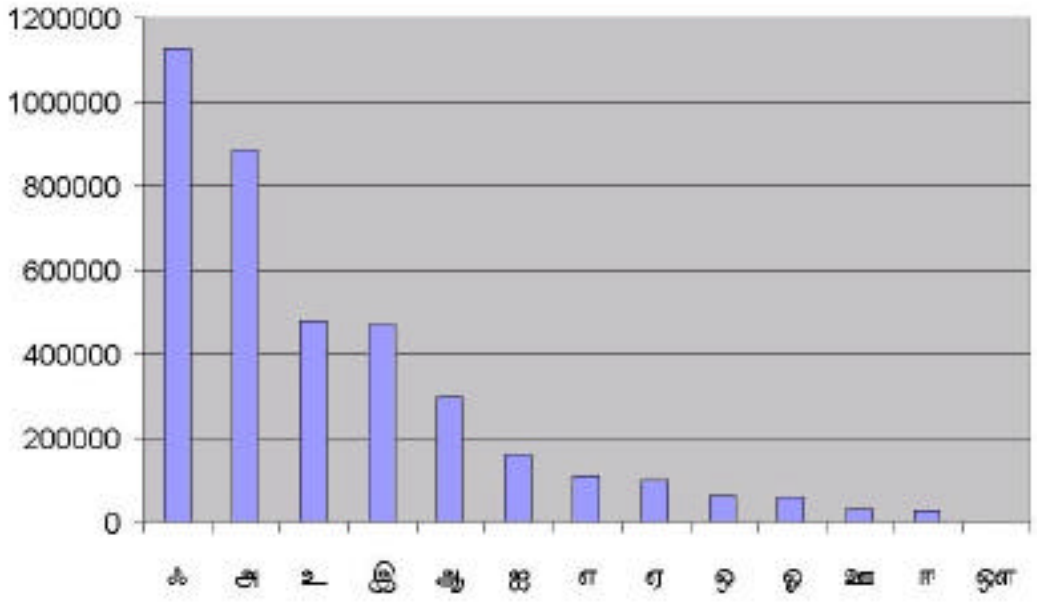

Figure 2. Frequency of Vowel Occurrences

Table 1. Consonants Keystrokes and Occurrences

Consonant	Keystrokes
த	491301
க	470502
ர	269817
ம	259966
ன	259648
ப	259610
ட	239337
வ	220345
ல	197443
ய	178709
ற	147874
ந	144849
ச	135136
ள	122274
ண	71821
ங	38102
ழ	32141
ஞ	7207

Table 2. Vowel Occurrences

Vowels	Occurrences
ஃ	1129347
அ	887947
உ	478840
இ	472581
ஆ	298354
ஐ	161061
எ	110330
ஏ	99559
ஒ	64255
ஓ	61162
ஊ	29956
ஈ	28390
ஔ	35

The total number of keystrokes to type the aggregation of works was calculated using the data for number of occurrences of each letter. Rule 4 of the Tamilnet99 Keystroke Sequences was taken into account. The data show an overall 1.2% increase (range ñ0.8% to 2.6%) in keystrokes for the ëalternateí keyboard (Table 3).

Table 3. Number of Keystrokes by Keyboard

Literary Selection	Tamilnet99 Keyboard	'Alternate Keyboard'	Percentage Difference
PM058	196430	198062	-0.8
PM065	65540	64603	1.4
PM066	29831	29070	2.6
PM082	105457	104585	0.8
PM088	102884	102172	0.7
PM097	11641	11530	1.0
PM098	54521	54149	0.7
PM099	17192	17079	0.7
PS	3165372	3129516	1.1
SS	1033276	1027887	0.5
AO	1269056	1241578	2.2
PK	261092	258738	0.9
Total	6312292	6238969	1.2

Discussion

The most obvious flaw in the Tamilnet99 keyboard design is the placement of the letter க. The data clearly show that க along with த is one of the two most widely-used consonants in the alphabet. However, க is placed at the 'h' key, a much weaker key position than the ëjí or ëkí keys, or even the ëlí or ë;í keys. Other individual vowel and consonant letters are also sub-optimally positioned on the keyboard.

The data on total number of keystrokes raise an important question about the decision to represent all consonants on the Tamilnet99 keyboard as அகர உயிர்மெய் letters (க, ங, ச, ..., ற, ...), as opposed to the 'alternate keyboard' with மெய் consonants (க், ங், ச், ..., ற், ன்). By using மெய் consonants in an 'alternate keyboard' arrangement, the number of keystrokes is increased by only a small percentage.

However, the representation of consonants by ¶¡ˆ letters results in a simpler learning process and avoids potential confusion associated with the Tamilnet99 Keystroke Sequences. Rule 4 of this annexure will almost certainly result in confusion for the learner when typing a combination of words such as "விரைவாகக் கிளம்பினாள்" Using the Tamilnet99 keyboard, the natural tendency of the user is to start typing the underlined segment as க + க + ஃ, but this will automatically get converted by Keystroke Sequence Rule #4 to "க்க்" instead of "கக்".

It is our opinion that Keystroke Sequences for a phonetic keyboard should be intuitive, effortless, and universally applicable. The 'alternate keyboard' has the potential to meet this goal more easily than the Tamilnet99 keyboard. To require a typist to perform mental

gymnastics and deduce appropriate keystroke sequence for certain letter combinations not only impairs typing speed, it indicates a design flaw.

By eliminating these confusing keystroke sequences and using an 'alternate keyboard' with மெய் consonants and no special keystroke sequences, we can simplify the phonetic keyboard and make the process of typing a more universally valid one. This will in turn increase the speed of adoption of the Thamil phonetic keyboard among world Thamils.

Conclusion

The data from this study clearly demonstrate sub-optimal key placement for several consonants and vowels of the Tamilnet99 phonetic keyboard. Before efforts to promote worldwide adoption of a Thamil phonetic keyboard are accelerated, we strongly advise small revisions of the Tamilnet99 keyboard to create a more-optimized keyboard standard. Specifically, the position of the க key should be changed to the strongest key positions, namely those keys corresponding to the English ëkí or ëjí key.

Also, we encourage Thamil computing bodies to consider a possible transition to an 'alternate keyboard' arrangement. We feel the special Keyboard Sequences are not always intuitive and may confuse the new learner. Moreover, creation of உயிர்மெய் letters through though an 'alternate keyboard' may be more intuitive.

While we understand and acknowledge the inconvenience caused to keyboard and software manufacturers by a keyboard redesign, we feel the long-term benefits to future generations of Thamil phonetic keyboard learners far outweigh the short-term cost.

Bibliography

Tamil 99 Keyboard Standard, Tamil Nadu Government Order No. 17, Annexure I, II, 1999.

Tamil99 Computer Keyboard, Sinnathurai Srivas, min manjari, Vol. 1, 2004.

Selection and Standardization of Tamil Keyboard Layouts - Recommendations of Tamil Nadu Standardization Committee, S. Kuppuswami, P. Venkatesan, Department of Computer Science, Pondichery University, India.

Typewriting Behavior, August Dvorak, Nellie L. Merrick, William L. Dealey, and Gertrude Catherine Ford, American Book Company (New York), 1936.

The Standard and Dvorak Keyboards Revisited: Direct Measures of Speed, Leonard J. West, Santa Fe Institute Working Paper, 1998.

Tamil Character Frequency Tables, Thuraiappah Vaseeharan. (All primary data and source code are available at this website.)