The Design of Khmer Word-based Predictive Non-QWERTY Soft Keyboard for Stylus-based Devices

Phavy Ouk[†], Ye Kyaw Thu*, Mitsuji Matsumoto[‡] and Yoshiyori Urano[¥]
Graduate School of Global Information and Telecommunication Studies, Waseda University
E-mail: [†]*ouk_phavy@fuji.waseda.jp, *ykt@akane.waseda.jp,*
[‡]*mmatsumoto@waseda.jp,* [¥]*urano@waseda.jp*

Abstract

We introduce a first ever soft keyboard for stylus-based devices such as Personal Digital Assistant (PDA) for Khmer, official language of Cambodia. The contribution of this study is twofold – a key layout arrangement and a word-based predictive text entry method. First, we design a non-QWERTY key layout in which consonants and vowels are grouped phonetically and orthographically, respectively. As a result, the number of soft keys is much less than that of Khmer character sets. Second, we present a word-based predictive text entry method based on the careful analysis of the structure of Khmer word composition. In spite of the word-based predictive mechanism, this soft keyboard provides as well the ability to input unknown words without swapping to other modes. A prototype has been developed and preliminary experiment shows that the proposed soft keyboard is user-friendly easy to use with little training. In addition to the application to stylus-based devices, it can also be extended to be a soft keyboard for conventional desktop.

1. Introduction

Computer software and modern devices such as mobile handsets available in Cambodia, one of Southeast Asia countries, have been mostly in English. This partially limits the full usage of hose devices by local people as not all of them are able to read and write in English. This limitation along with the idea of language localization of the country utterly encourages the application of Khmer language to those devices. Up to now, some of computer software developed by KhmerOS (a project team for the implementation of open source software in Cambodia) [1] and a certain kind of mobile handsets [2] support Khmer language. However, to the best of our knowledge, currently there is no Personal Digital Assistant (PDA) launched in Cambodia supports Khmer language yet. Seeing current soft keyboard features QWERTY keyboard layout, we conjecture that future PDA soft keyboard layout for Khmer may also adopt the QWERTY keyboard. When the large syllabic character sets of Khmer are mapped on the QWERTY keyboard, each key is assigned to many letters, 2 or 3, resulting in many mode usages requiring users to memorize the location of each character on the keyboard.

Therefore, this paper introduces a first ever word-based predictive non-QWERTY soft keyboard for stylus-based devices for Khmer language. Our proposed soft keyboard is easy to learn, demands less memory, and uses fewer keys in comparison to the current conventional desktop soft keyboard. Furthermore, preliminary evaluation provides an encouraging result of better typing speed of approximately 15% and comparable keystrokes.

2. Khmer language

2.1. Introduction to Khmer language

Cambodian, also known as Khmer, which is the official language of Cambodia belongs to the Mon-Khmer group of Austro-Asiatic languages and has been considerably influenced by Sanskrit and Pali. Khmer differs from neighboring languages such as Thai, Lao and Vietnamese in that it is not a tonal language. The writing system begins on the top left of the page, and proceeds down and to the right, with characters being placed above, below the main line of writing, leaving no space between words. A space in Khmer is a punctuation sign similar to a comma. Khmer contains 33 consonants, 32 subscript consonants with a pair of duplicates, 24 dependent

1-4244-2528-0/08/$25.00 ©2008 IEEE

vowels, 12 independent vowels, 2 consonant shifters, and a dozen diacritic signs and other symbols including number signs [3]. Most consonants have reduced or modified forms, called subscripts, when they occur as the second member of a consonant cluster. Figure 1 demonstrates Khmer Character symbols.

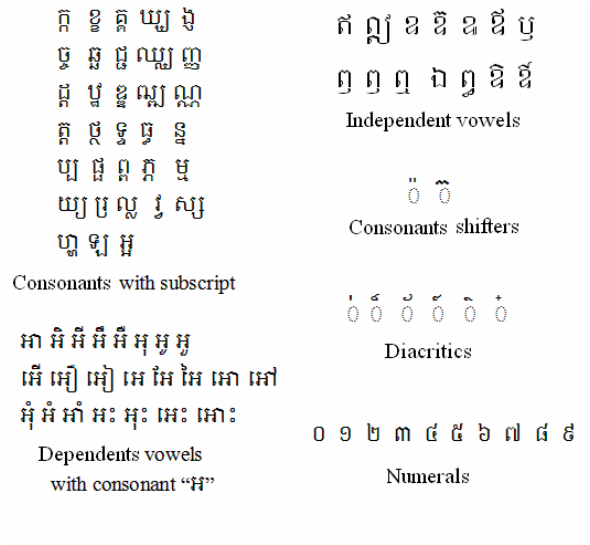


Figure 1. Khmer character symbols

2.2. Khmer Unicode keyboard layout

With the past typing system, Khmer word is tapped on typewriters and computers from left to right. Moreover, subscripts, small form of consonants, are spread over the keyboard. A new keyboard based on Khmer Unicode developed by KhmerOS (Khmer Software Initiative) [4], brings a different way of writing — the typing order is the same as spelling order, not handwriting order which is from left to right. Furthermore, subscripts of the consonants are not encoded on the keyboard anymore. Instead, a subscript sign "◌្" is used to indicate that the next consonant acts as a subscript. Featuring the QWERTY keyboard layout where Khmer characters are mapped according to the phonetic Romanization corresponding to the alphabetic character, Khmer keyboard is forced to employ shift mode and right-alt mode in addition to normal mode. This is because Khmer comprises of so many character sets. Therefore, typists need to hold down many keys simultaneously when typing the letter that is not in normal mode. Figure 2 shows the letter arrangement of the keys of the current Khmer Unicode Keyboard.

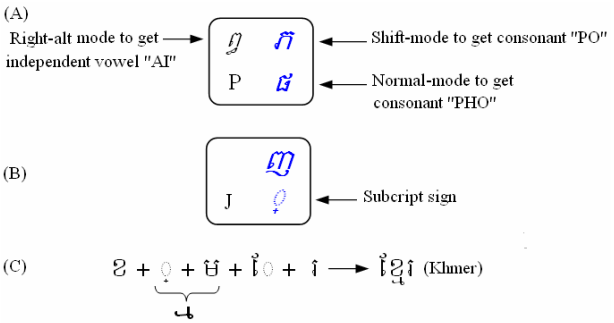


Figure 2. (A) The letter arrangement on a key of current Khmer Unicode keyboard, (B) Subscript sign on the keyboard, (C) Typing order of the word ខ្មែរ (Khmer) by using Khmer Unicode Keyboard

2.3. Khmer soft keyboard

Khmer Unicode soft keyboard adopting the Khmer Unicode keyboard layout is developed by Tavultesoft [5]. As mentioned in the preceding section, Khmer Unicode keyboard uses many modes—normal, shift and right-alt. When working with conventional desktop keyboard, users can concurrently force down two keys to get a letter that is in shift mode or right-alt mode. However, with soft or virtual keyboard this action can not occur because users can not have two mice to click two keys at the same time. Thereby, every time a shift or right-alt mode character is needed, shift key or right-alt key needs to be firstly pressed before getting the target character. Figure 3, 4 and 5 depict Khmer soft keyboard layout in normal mode, shift mode, and right-alt mode, respectively.

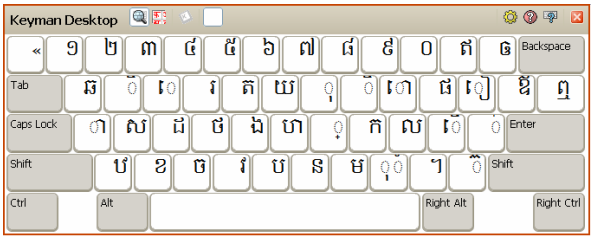


Figure 3. Khmer soft keyboard in normal mode

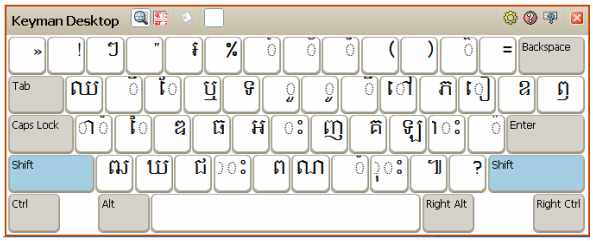


Figure 4. Khmer soft keyboard in shift mode

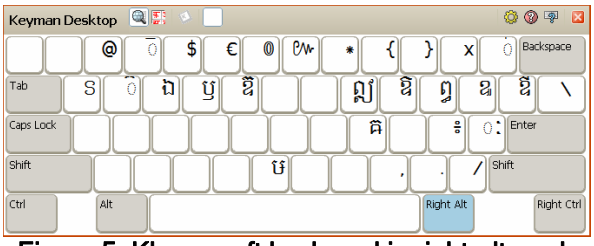


Figure 5. Khmer soft keyboard in right-alt mode

3. Proposed soft keyboard

The work has two contributions, key layout arrangement and predictive text entry, which will be disclosed in the following sections.

3.1. Key layout arrangement

After a careful study on the sound of the consonants and the shape of the vowels, we arrive at the following key arrangement:

3.1.1. Consonant mapping. Khmer Consonants are divided into two voices. One of them is called 'Akhosak', also known as voiceless or first series. The other is called 'Khosak' or voiced or second series. In Table 1, for the audio reference the voiceless series is depicted in italic and the voiced series in regular. Taking this feature of Khmer language into consideration, we classify the consonants into two, first series group and second series group, which are placed at the upper part and the lower part, respectively (Figure 10). This key layout is naturally arranged in the order of the Khmer consonants, except the last row consonants where originally the second series comes first. We refer readers to Figure 1 for the original order of Khmer consonants. Figure 6 exemplifies the first row consonants in the original order and their re-arrangement in the proposed interface.

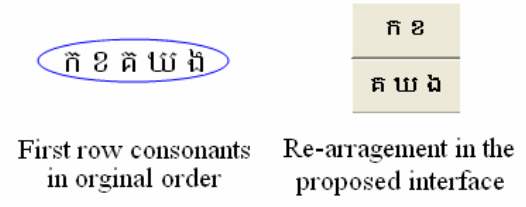


Figure 6. The first row consonants in original order and in the proposed interface

3.1.2. Vowel and diacritic mapping. We categorize Khmer dependent vowels into 3 distinct groups according to its shape as follows:

1. ◌ា, ◌ិ,◌ី,◌ឹ, ◌ឺ, ◌ុ,◌ូ,◌ួ: "◌ា" or "A" group
2. ◌ើ, ◌ឿ, ◌ៀ, ◌េ,◌ែ,◌ៃ,◌ោ,◌ៅ: "◌េ" or "E" group
3. ◌ំ,◌ុំ,◌ាំ,◌ះ, ◌ុះ, ◌ូះ, ◌ោះ: "◌ំ" or "O" group

Herein we give the priority to "◌ំ" group or "O" group meaning at the first sight of considering the target vowel users need to visualize firstly the "◌ំ" sign. If it exists, they just go directly to this group ("O" group). If the target character does not contain "◌ំ" sign, users need to scan for "◌េ" sign, and then go to "E" group if it does contain the sign. "A" group is the last home to be in, if the desired vowel is composed of neither "◌ំ" sign, nor "◌េ" sign.

Each of above group is categorized into 2 subgroups:

1. "A" group: (◌ា,◌ិ,◌ី,◌ឹ,◌ឺ) (Up vowels)

   (◌ុ,◌ូ,◌ួ) (Down vowels)

2. "E" group: (◌ើ, ◌េ,◌ែ,◌ៃ) (Right-up vowels)

   (◌ោ,◌ៅ, ◌ឿ, ◌ៀ) (In between vowels)

3. "O" group : (◌ំ,◌ុំ,◌ាំ) ("◌ំ" sign vowels)

   (◌ះ, ◌ុះ, ◌ូះ, ◌ោះ) ("◌ះ" sign vowels)

Table 1 shows the key layout of the proposed interface.

Table 1. Consonant, vowel and diacritic mapping

Consonant	ក	ខ	គ	ឃ	ង			
	ច	ឆ	ជ	ឈ	ញ			
	ដ	ឋ	ឌ	ឍ	ណ			
	ត	ថ	ទ	ធ	ន			
	ប	ផ	ព	ភ	ម			
	ស	ហ	ឡ	អ	យ	រ	ល	វ
"A" group	◌ា	◌ិ	◌ី	◌ឹ	◌ឺ	◌ុ	◌ូ	◌ួ
"E" group	◌េ	◌ែ	◌ៃ	◌ើ	◌ឿ	◌ៀ	◌ោ	◌ៅ
"O" group	◌ុំ	◌ំ	◌ាំ	◌ះ	◌ុះ	◌ូះ	◌ោះ	
DS, CS	◌់	◌៉	◌៊	◌៌	◌័	◌៎	◌៏	◌៑
In.V, S	ឥ,ឦ,ឧ,ឩ,ឪ,ឫ,...etc				។,៕,៖,ៗ,๏,ฯ,...etc			

DS: Diacritic signs; CS: Consonant shifters
In.V: Independent vowels; S: Special signs

3.2. Word-based predictive concept

Predictive text entry was created to overcome the problem where too many characters need to be assigned to the limited number of keys on a keyboard or a mobile handset. The predictive text entry starts when a correct order character sequence of the target word is input, which then is constructed and compared to the words in an embedded dictionary in order to search for the corresponding target candidates and display them to users, and a selection of the target word is the final stage [6]. The method is able to produce corresponding target candidates if the character sequence of the target word is input correctly. If any single character of the target word is missing, the system will end up with words that are not expected.

Our word-based predictive method also enlists help of dictionary of known words which is used to compare with a character sequence of the target word. However, with our proposed scheme, the subscript sign is excluded from the character sequence of the target word, which means that when constructing a word, users can neglect subscript sign by just going straight to the main consonant. To simplify the understanding, we will give an overview of Khmer word typing order using current Khmer Unicode keyboard. As mentioned in earlier section, in current Khmer Unicode keyboard and soft keyboard, in order to type a subscript of a consonant, two characters are needed—a subscript sign and a main consonant of the target subscript. Figure 7 illustrates the name of each character symbol of the word "ស្ត្រី" (woman) and subscript sign. Repeatedly mentioned, with our proposed text entry, subscript sign is omitted, which means that users do not have to input subscript sign anymore when they need a subscript of a consonant, instead they just input the main consonant of the target subscript. Table 2 depicts the difference of typing order of the word "ស្ត្រី" (woman) of conventional typing and proposed method's typing.

In addition to character reduction typing, our approach guesses 2 characters ahead of current character sequence of the target word. This means that the system includes in the candidate list both words that correspond to the character sequence of the target word and words that have two characters longer than the character sequence of the target word. The idea of two-character ahead guessing does not come by chance; we have a look at today's text typing of Khmer Unicode structure. In order to display a subscript of a consonant, a subscript sign and a main consonant character are required, which conveys the meaning that to display a subscript two characters are needed. Moreover, we conducted a pilot scheme on guessing the rest of the characters starting from the current input character sequence, and the outcome is that there are too many words displaying on the candidate list, which makes the system slow and confuses users since they spend much time for a single selection from those hundred suggested words.

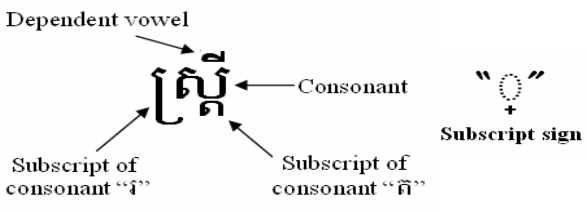


Figure 7. Name of each character symbol of the word ស្ត្រី (woman)

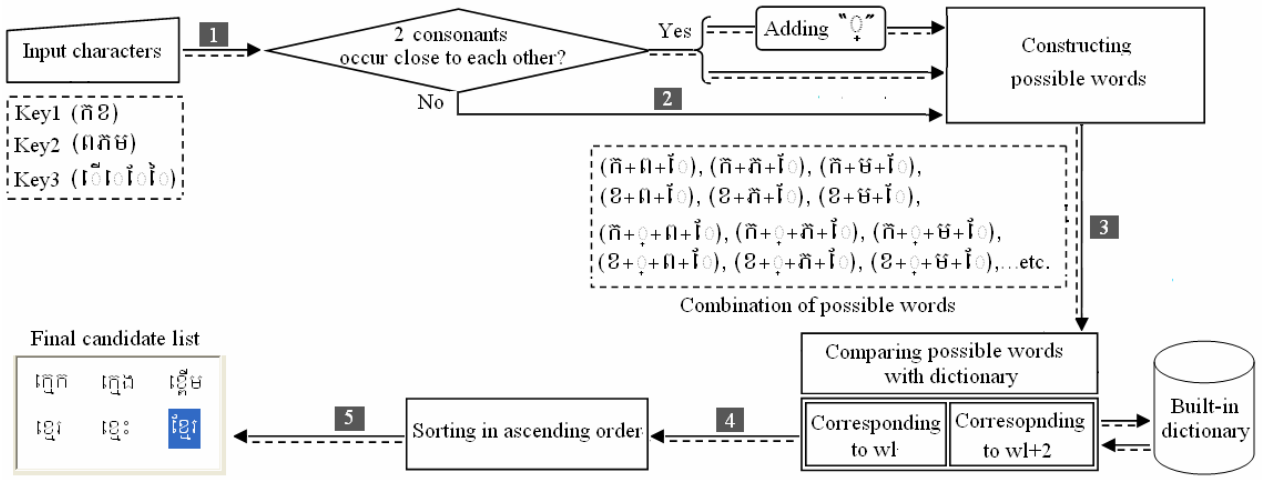
Table 2. Different typing order of current Khmer Unicode Typing and proposed method

Input Method	Word	Typing Order
Khmer Unicode keyboard	ស្ត្រី	ស + ◌្ + រ + ◌្ + ត + ◌ី
Proposed method	ស្ត្រី	ស + រ + ត + ◌ី

3.3. Process flow of proposed method

The architecture of our method is demonstrated in Figure 8.

1. The system begins when users input a character sequence of their target word.
2. The above character sequence is checked whether two consonants appear close to each other. If it does, the system will combine the two sequence consonants in two different ways – 1, the system will directly combine the two consonants and 2, a subscript sign "◌្" will be embedded between the two consonants. However, if the character sequence composes of a consonant followed by a vowel, the system will directly combine those characters without adding anything and proceed to the next step.
3. Possible words are constructed from the input character sequence combined with the embedded subscript sign if it was added in the previous process.



wl: length of words in built-in dictionary
Possible words: the combinations of the character sequence input by users

Figure 8. Process flow of proposed interface

4. A built-in dictionary is used to compare to the possible words constructed in step 3. The length of possible words is also checked and compared with the length of the words in dictionary. The system extracts the dictionary words that match the length of the possible words, and words that have 2 characters longer than the possible words. Those words then are added to a potential candidate list.
5. The candidate list is arranged in ascending order, so that the user can first access to the word that match to the exact character sequence. A word list is displayed to users and a choice from that list is the final step of this process.

For a clear understanding to our system's concept, we will provide an example of typing the word ខ្ំមរ (Khmer). The word comprises of 5 characters, which are "ខ", "្", "ម", "ំ◌", and "រ" (This is the typing order using Khmer Unicode Keyboard). The dashed line in Figure 8 illustrates the flow of the system to generate the example word. First, because our system reduces a subscript sign and can guess 2 characters ahead of the current typing characters, users press only 3 keys (កខ), (ពភម), and (ើ◌ៈ◌ំ◌ៃ◌ុ◌ូ◌ោ◌ៅ), respectively. Second, as two consonants are input after one another, the system combines all of the input characters in two different ways, 1- with subscript sign embedded and 2- no subscript sign. Third, the system constructs the possible words: (ក+ព+ំ◌), (ក+ភ+ំ◌), (ក+ម+ំ◌), (ខ+ព+ំ◌), (ខ+ភ+ំ◌), (ខ+ម+ំ◌), (ក+្+ព+ំ◌), (ក+្+ភ+ំ◌), (ក+្+ម+ំ◌), (ខ+្+ព+ំ◌), (ខ+្+ភ+ំ◌), (ខ+្+ម+ំ◌),...etc. Due to the space limitation, we only show the combination with a vowel (ំ◌), instead of all vowels (ើ◌ៈ◌ំ◌ៃ◌ុ◌ូ◌ោ◌ៅ). Fourth, those combinations will be compared to the built-in dictionary to find the words that correspond to the length of the input character sequence and that have two characters longer than the input character sequence. Fifth, a candidate list will be displayed to users: "ក្មក", "ក្មង", "ខ្ចីម", "ខ្មរ", "ខ្មុះ", "ខ្ំមរ", after sorting in ascending order. In fact, users can type all of the characters used to compose the target word, if they want to shorten the candidate list. In this case, if users press another key (យ រ ល ំ) for "រ" character they will have a candidate list of only two words which are "ខ្មរ", and "ខ្ំមរ".

4. Experiment and evaluation

4.1. Subjects and Procedure

We recruited five volunteer subjects; all of them are Cambodians with the age ranging from 24 to 28. Subjects were trained the basic use of Khmer soft keyboard and the proposed soft keyboard before starting the experiment, and thus to familiarize themselves with the methods, they were permitted to

key in either the practice words or their own words, which took about 20 minutes. Subjects were told that their typing speeds will be recorded, yet they must correct all of the incorrect typing as many as possible. Subjects were asked to type 5 Khmer sentences used in daily conversation between friends, and composed of most of the consonants, vowels, numbers, and diacritic signs (Figure 9). We tracked the users' time to complete the whole dialog which each user was requested to type 5 trials. The prototype of the proposed interface (Figure 10) is developed with Visual Basic. Net [7], and uses on Chuon Nath dictionary [8] in the current Unicode format as the built-in dictionary.

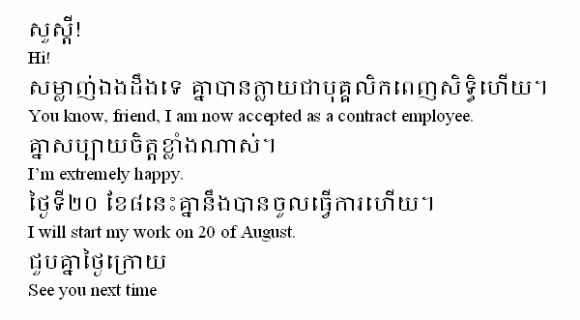
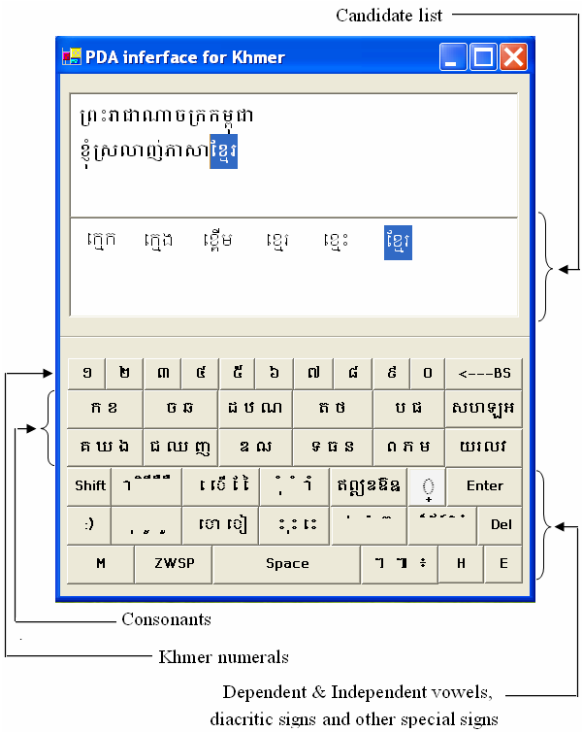


Figure 9. Experiment sentences



:) :Smiley; M: Mode (change to English interface)
ZWSP: Zero width space(for invisible word break)
Figure 10. Interface of proposed soft keyboard

4.2. Speed (Characters per minute)

Generally, words per minute (WPM) is mostly calculated to report the speed of a text entry method in similar papers. The common definition for "word" is a term of 5 characters including space [9]. However, in our case, it's not reliable if we focus on WPM as the number of characters per word of syllabic script and that of alphabetic language might not be the same. We hesitate to measure 5 characters per Khmer word as there is no exact source. Therefore, we based on characters per minute (CPM) to evaluate the performance of the proposed method. Herein, we calculate CPM by dividing the transcribed text which consists of 135 characters with the completion time of each trial of user in minutes.

We have three categories of subjects — First category refers to the subject that experiences neither Khmer Unicode keyboard nor Predictive method (1 user). Second category is the subject that has experience Khmer Unicode keyboard typing, but not predictive method (1 user). The subject types at least 20 Khmer sentences per week. The third category consists of 3 subjects who have basic knowledge of both Khmer Unicode keyboard and Predictive method. They type 6 Khmer sentences per week, and use predictive method in English or Chinese to type short messages of about 5 sentences per week. We will have a discussion on each category as follows:

Figure 11 depicts CPM of the first category subject. At the first trial, subject typed 21 characters per minute with proposed method and 14 CPM with existing method. The speed of both methods gradually grows in parallel. The result shows that our method is easier to learn and to adapt because by providing the same practice time, subject can type faster with our method.

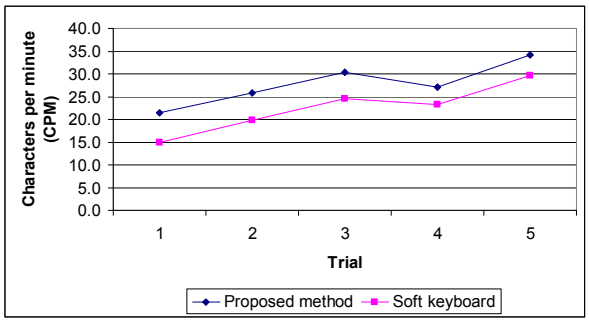


Figure 11. CPM of 1 subject with no experience of both methods

The second category subject started with 20 CPM with Khmer soft keyboard and 23 CPM with proposed

method. However, at the last trial, the speed of the two methods is comparable (Figure 12).

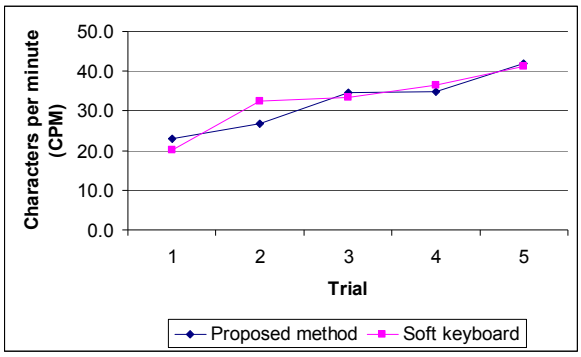


Figure 12. CPM of 1 subject with experience of Khmer Unicode typing but not of predictive method

Average speed comparison of the two methods of the third category subject is shown in Figure 13, where CPM of the proposed method is higher in each trial in comparison to that of the existing one. We observe that the graph of the third category subject and that of the first category subject yield similar trend that CPM of the proposed method is higher than that of the existing method. This is telling us that our method is more user-friendly since the subjects that have the same experience of both methods could go faster with our proposed method even at the first time.

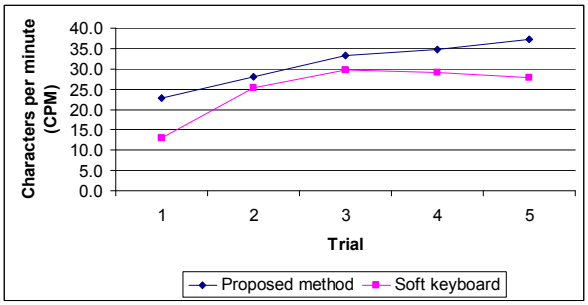


Figure 13. CPM of 3 subjects with experience of both methods

In average, our proposed soft keyboard (31 CPM) yields better speed of 15% compared to the conventional soft keyboard (26 CPM).

4.3. Keystrokes per character (KSPC)

We calculate keystrokes per character by using an evaluation metric of Mackenzie [10]. We examine the experiment sentences which consist of 135 characters. As a whole, there is very little difference in terms of KSPC between the proposed method and Khmer soft keyboard. As shown in Table 3, Khmer soft keyboard needs 1.22 keystrokes for a character while our method consumes 1.19. At first glance, the proposed method seems to reduce KSPC compared to the current Khmer soft keyboard as it reduces a character "subscript sign." The reason leads to comparable keystroke is that the proposed method needs an extra key-press to choose the desired character from the candidate list. There is the case that users don't have to use the extra key; it is when the candidate list contained only 1 suggested word.

Table 3. Keystrokes per character (KSPC) of Khmer soft keyboard and proposed method

Input Method	KSPC
Proposed method	1.19
Khmer soft keyboard	1.22

4.4. Discussion

The user study shows that our proposed soft keyboard yielded better speed of 15% in overall compared to conventional soft keyboard and comparable keystrokes though the method uses much fewer keys. We have tested on three category subjects; the experiment provided such an encouraging result that all category subjects started faster with our proposed interface. We have also found a hindering problem that makes the speed of the proposed interface slow. It is that subjects have less experience with predictive idea; some have basic knowledge with predictive method in English or Chinese, and others know nothing about the predictive concept. Specifically, it is their first time to practice word-based predictive method in Khmer language. From the experiment we learned that it took time to think of the character order to input their target words. They don't get used to the habit of spelling the word before typing, but they mostly visualize the words, as they used to do in handwriting.

Furthermore, we were reported by subjects that the key arrangement of the proposed soft keyboard is very easy to get familiar with, and that both consonants and vowels are well arranged in natural order with a clear guideline; they thereby do not have to memorize the position of the target character and look around to find it. Therefore, due to our user-friendly key layout we strongly believe that providing more training time on word-based predictive concept with Khmer language to users will yield better speed.

4.5. Merit of the proposed method

- The key layout does not employ many modes in sprite of the large character sets of Khmer language. Thereby, users do not have to memorize which character is in which key and at which mode.
- Consonants and vowels are naturally grouped and placed on the soft keyboard, thus users spend less time to scan for their desired character.
- Less number of keys on the soft keyboard provides the bigger size of each key when displaying on the small screen of PDA, users thereby can access to those keys with their finger.
- We shorten the number of character sequence of the target word by eliminating the subscript sign "្" resulting in less complicated writing.
- Not only dictionary based approach, the system also allows users to write unknown words without having to switch to other modes, as we embed a subscript sign key on the soft keyboard.

4.6. Critique of the method

The experiment was conducted with only a limited number of users and for a short period, 5 trials of the same sentences, and did not address the long term use. A long term study with more subjects would prove the better performance of the proposed interface in terms of speed.

The experiment sentences are commonly used in text message by friends and hence they are not representative of all character of Khmer language.

The prototype still provides a quite long candidate list, which slows down users' selection of their target word. Actually, this can be solved by increasing the number of keys on the soft keyboard. The more keys on the keyboard, the fewer the character symbole on each key. However, doing so the key arrangment may not be in natural order as it is now, which will take more time for key finding. Morever, as modern devices become smller, it is not a good solution to add more keys on the interface. To avoid this trade-off, we instead are thinking of sorting the candidate list according to frequency used words or users' history for better speed. We plan to reach this goal in the near future.

As reported by subjects, the word-based predictive concept in Khmer is very new for them, and thus they find it hard to visualize the character order of the target word and type at the same time. It is becaue they have the habit of writing from left to right in handwriting. Therefore, more training on Khmer word-based prediction needs to be provided, especially to those who have no experience at all with predictive method.

5. Conclusion and future work

We have described a non-QWERTY soft keyboard for PDA in Khmer language. Our contributions are a key layout arrangement and a word-based predictive method for Khmer. We use a very limited number of keys on the soft keyboard, and sub-group the consonants and vowels in a natural way resulting in less time-consuming and human memory-consuming. Moreover, we shorten the number of characters of the target word resulting in less complicated writing. The experiment shows that the speed of our method is 15% faster than that of the current soft keyboard. We also observe that our system is user-friendly in terms of key layout and can be learnt with a very little training time. As our future work, we will focus more on the word frequency as well as the user's own corpus based on recent context. We believe that by eliminating the uncommon word candidates, the text input would be more efficient.

References

1. "Khmeros," http://www.khmeros.info/drupal/
2. "Mobile phones support Khmer language," http://kth.com.kh/index.php
3. Huffman, F. E. 1970. Cambodian System of Writing and beginning reader with Drills and Glossary. Yale University Press.
4. Link to download document on Khmer Unicode font and how to type Khmer Unicode: http://www.khmeros.info/drupal/?q=en/download/docs
5. Link to download Khmer Unicode soft keyboard layout using the NiDA keyboard layout http://www.tavultesoft.com/keyman/downloads/keyboards/details.php?KeyboardID=401
6. Tanaka-ishii, K. 2007. Word-based predictive text entry using adaptive language models. *Nat. Lang. Eng.* 13, 1 (Mar. 2007), pp. 51-74.
7. "Visual Basic.Net," http://www.microsoft.com/express/vb/
8. Chuon Nath. 1967. Dictionnaire Cambodgien. L'INSTITUT BOUDDHIQUE
9. Yamada, H. 1980. A historical study of typewriters and typing methods: From the position of planning Japanese parallels. Journal of Information Processing, 2, 175-202
10. Silfverberg, M., MacKenzie, I. S., & Korhonen, P. 2000. Predicting text entry speeds on mobile phones. Proceedings of the ACM Conference on Human Factor in Computing Systems—CHI2000, pp. 9-16. New York: ACM