Journal of Computer Science 1 (2): 189-199, 2005
ISSN 1549-3636

Investigating Text Input Methods for Mobile Phones

Barry O'Riordan, Kevin Curran and Derek Woods
School of Computing and Intelligent Systems
University of Ulster, Magee Campus, Northland Road, Northern Ireland UK

Abstract: Human Computer Interaction is a primary factor in the success or failure of any device but if an objective view is taken of the current mobile phone market you would be forgiven for thinking usability was secondary to aesthetics. Many phone manufacturers modify the design of phones to be different than the competition and to target fashion trends, usually at the expense of usability and performance. There is a lack of awareness among many buyers of the usability of the device they are purchasing and the disposability of modern technology is an effect rather than a cause of this. Designing new text entry methods for mobile devices can be expensive and labour-intensive. The assessment and comparison of a new text entry method with current methods is a necessary part of the design process. The best way to do this is through an empirical evaluation. The aim of the study was to establish which mobile phone text input method best suits the requirements of a select group of target users. This study used a diverse range of users to compare devices that are in everyday use by most of the adult population. The proliferation of the devices is as yet unmatched by the study of their application and the consideration of their user friendliness.

Key words: Mobile Human Computer Interaction, Input Text Methods

INTRODUCTION

The ability to record information on a retrievable medium and to recall it for later use has been in existence in one form or another since the first stone-age etchings. The development of methods and media has progressed through Egyptian hieroglyphics on stone, ink on parchment, type on paper, to pixels on liquid crystal displays. The second half of the last century has seen the majority of advances in this area. The method of text input has changed little, however, in the same period. The initial QWERTY keyboard won out over its closest rival, the more efficient DVORAK keyboard, not through superiority, but through politics. The QWERTY keyboard has become the international standard and such widespread use has made change virtually impossible. The retraining time and replacement cost are two factors that prevent more efficient text entry methods from being adopted. The QWERTY keyboard by design was intended to slow the typist down in order to prevent the print arms of the old typewriter from jamming at the point where they make contact with the paper. Though technology has overcome this mechanical problem, the keyboard has remained unchanged and so text input speed has not improved beyond that of the early days of typing [1]. The devices that are most commonly used for entering text are personal organisers or Personal Digital Assistants (PDA's) and mobile phones. The most common methods for text entry on these devices are the ten digit number pad with letters assigned to different numbers, a miniaturised 'hard' or 'soft' QWERTY keyboard, or handwriting recognition using a stylus on an LCD touch screen. Companies trying to sell the virtues of their new text input methods have done numerous studies, but few studies compare the existing methods from a user perspective [2, 3]. Portability and the desire to have information immediately to hand have caused a rapid growth in the PDA market. With it, the need for a suitable text entry method has become apparent. This study concentrates on the following devices as they represent a broad spread of devices in common use among the general public:

* Computer Keyboard (QWERTY)
* Personal Organiser (miniature QWERTY keyboard)
* Mobile Phone keypad
* PDA soft QWERTY keyboard
* PDA handwriting recognition

Little work had been done on how people perform when entering text for personal use. Fitts' Law for predicting the speed of text input methods is overly relied upon as being the definitive standard for assessment [4]. Time spent in real testing is generally deemed time wasted. This study goes back to the basic method of user testing because, as the study will show, there are as many variables as there are people.

The choice of phones was such that there was a 'standard' keypad-Nokia 6310 (Fig. 3), a 'non-standard' keypad Nokia 7210 (Fig. 2) and a Motorola Accompli 008 (Fig. 4) which doubled up as the PDA soft QWERTY and PDA Handwriting devices. The other devices chosen were a Compaq Armada E500 laptop with a standard computer keyboard attached.

This was chosen to establish a base line of typing ability as well as to compare the effect of size reduction when going from a full size QWERTY keyboard to a miniaturised QWERTY keyboard, as on the Sharp Organiser (Fig. 1) and Motorola PDA soft keyboard.

Fig. 1: Organiser

Fig. 2: Nokia 7210

Fig. 3: Nokia 6310

Fig. 4: Accompli 008

Related Work: A number of empirical studies have been done in the area of text input in mobile devices, [2, 3, 5, 6]. Dunlop *et al.* [5] looked at the development of word prediction, which is a logical progression from word recognition. This method would allow the user to input fewer letters than comprised the word by selecting the desired word from a list of suggestions which would reduce with the further completion of the inputted word. McKenzie [6] examined the *LetterWise* system which predicts the next letter to be entered when inputting words. They compare LetterWise to the traditional Multitap or non-predictive text method. James [3] was concerned with the performance of Predictive (T9) and Non-predictive (Multitap) text input methods when compared to predictions from two different mathematical models. Their study was similar in that results were presented in terms of accuracy and speed of text input for users of different experience levels. We differ in terms of text input interface and style of text inputted. Butts and Cockburn [2] presents an empirical study that compares three mobile phone text input techniques. They are 'multi-press input with timeout', 'multi-press input with a next button' and two-key. Again, this study concentrates on the traditional phone keypad for text input.

Selection of users is a prime consideration in these types of experiments. The number of subjects performing the experiment of previous authors varied widely with no real attempt made to get a statistically acceptable sample size. The type of experiments being performed did not always lend themselves to large sample sizes due to the complexity of the devices being examined. Isokoski and Raisamo [7] used only five subjects, two female and three male university staff members aged 23 to 29 years. This was due to the prolonged training sessions required to test a new device for entering text. Butts and Cockburn [2] used eight subjects, all of whom were male postgraduate computer students. They were chosen such that three were novices, three were intermediates and two were experts. They defined Novices as those who never sent SMS text messages. Intermediates sent up to five SMS text messages per week. Experts sent more than five messages per week. Their study formed the basis for this dissertation and so their results should compare to those of this experiment though exact comparison will not be possible due to the different devices used in both studies and due to the different definitions of Novices, Intermediates and Experts. James and Reischel [3] used twenty subjects, ten male and ten female. The source of the subjects is not stated. They divided the group into Novices (0 SMS per week) and Intermediates (5+SMS per week). Only two of their subjects were considered to be experts sending more than ten SMS messages per week. Dunlop and Crossan [5] used fourteen subjects but no details are given of the subjects. The largest sample

size used was that of [8] which used twenty-eight subjects divided into four groups–Beginner (zero messages per week {2 subjects}), Novice (less than five messages per week {14 subjects}), Intermediate (five to fifteen messages per week {7 subjects}) and Expert (greater than fifteen messages per week {5 subjects}). All were either second or fourth year Computer Science students. All the studies mentioned, with possibly the exception of Dunlop and Crossan [5] used subjects educated to third level. This may skew as the sample is not a cross section of society. The spread of experience of the subjects was also uneven with unequal distribution of subjects within the various categories. Gender balance is considered by James and Reischel [3] however gender is not thought to be a significant factor but this can only be assessed by experimentation.

MATERIALS AND METHODS

The Nokia 6310 standard keypad phone and the Nokia 7210 non-standard keypad phone had the facility to enter text with or without predictive text. Predictive text is where a word is spelled by pressing only one key per letter. The phone has a large dictionary of words stored in its memory and selects a word from the in-built dictionary that matches the sequence of key presses. If the word displayed is not the intended word, a selection of choices can be scrolled through using, in the case of Nokia phones, the * key which is located under 7 on the keypad. Predictive text is defined as word level disambiguation where the system compares a sequence of ambiguous keystrokes to words in a large database to determine the intended word [11]. The non-predictive or multi-tap method is where the relevant key is pressed the corresponding number of times to get the letter that is located on the particular number key. In simple terms, to input 'C' the 2 key is pressed three times in quick succession. If another letter on the same key is required, a pause of between 0.5 and 1.0 sec. is required before the same key can be pressed again to get the second letter. The text to be entered was selected so as to examine the effect of different styles of phrases being entered on the different devices as shown in Fig. 5.

a. I have never sent a text message before
b. Your flying lesson's cancelled today. Call Andrew from 7:00 pm onwards to arrange another lesson.
c. Plane gets in at 10:00 pm to Gate 11. Aerlingus flight No. EI 987. Can you meet me? My e-mail address is * biggles@hotmail.com!
d. let me no where u r and il pic u up l8r

Fig. 5: Test Phrases

The first phrase is simple and allows the subject to enter text only without worrying about syntax or punctuation. The second phrase is moderate in complexity and requires the subject to use some punctuation and upper and lower case letters. Numbers are also introduced at this point. The third phrase is complex and tests the subjects ability to navigate the entire text input interface. The fourth phrase is in the style of an abbreviated message that uses common abbreviations for words and phrases. Designing new text entry methods for mobile devices can be expensive and labour-intensive [8, 9]. The assessment and comparison of a new text entry method with current methods is a necessary part of the design process. The best way to do this is through an empirical evaluation. Unfortunately, such evaluations are time-consuming and complicated. Careful planning and execution is needed when undertaking such an evaluation because an abundance of confounding factors exists that could negatively effect its repeatability and validity [8]. To control confounding factors, empirical evaluations place participants in constrained, artificial environments. This allows the behaviour or behaviours of interest to be isolated and thus accurately measured. To ensure the validity of an evaluation, however, it has to be designed to be as representative of actual user behaviour as possible [8, 9]. Sirisena [8] Suggest this need not result in a trade-off between accuracy and relevancy and evaluations should be designed to maximise both relevancy and accuracy.

Subject Group: Twenty-four subjects were selected. It was decided that there would be two of each category of subject in three experience levels, grouped by age and gender. This equated to two young males, two young females, two old males and two old females, in each of three categories; Novice, Intermediate and Expert. Subjects less than thirty years of age were considered young with those over thirty defined as old! Within this range the oldest subject was a fifty-two year old expert male and the youngest were thirteen-year-old novice male twins. Experience level was based on SMS message sending, though the questionnaire asked for an approximate Words-per-Minute for each subject on a standard QWERTY keyboard. Novices were those who sent less than five SMS messages per week, Intermediates sent between five and fifteen SMS messages per week and Experts were those who sent more than fifteen SMS messages per week. It was clear early in the conceptual stage of the experiment that finding subjects willing to do the experiment that had never sent an SMS message would be extremely difficult. The exponential growth in mobile phone usage has meant that the definition of user experience must be reviewed over time. What constituted an expert a few years ago would be re-classed as intermediate. The educational and professional backgrounds of the subjects varied from second level students (two males and two females), third level students or graduates (three males and eight females), white collar without third level education (three males), blue collar workers

without third level education (five males and one female). Subjects were also asked in the questionnaire if they were right or left-handed. The only notable difference during the experiment was that the PDA, when in the handwriting mode, did not recognise the letter "O" if drawn clockwise as is the general method used by left-handed people.

Test Devices: Within the traditional mobile phone layout category - the Nokia 6310 was selected. The Nokia 7210 was chosen for its non-standard keypad design. Both of these phones were used with the predictive text (T9) function and without the predictive function (referred to as Non-predictive, Multi-tap or Multi-press). A Motorola Accompli 008 stylus phone that also doubled as a PDA was tested using a stylus to input text both on a tiny 'soft' QWERTY keyboard and using handwriting recognition (no new alphabet needed to be learned). The other two devices tested were a full size QWERTY keyboard and a Sharp ZQ-4450 personal organiser that used a miniature QWERTY keyboard to input text.

Test Phrases: Test phrases (Fig. 5) were chosen with a number of expectations regarding the subjects' abilities to input them on each device.

Test Phrase 1: This phrase will present few problems to any subject and serve to ease them into the experiment. It will allow subjects to gain confidence with the form of input that they do not use on a daily basis, namely predictive or non-predictive.

Test Phrase 2: This phrase will cause only minor problems for novices and intermediates that are not accustomed to using punctuation marks and capital letters in the middle of sentences. Some experts will have the same problem.

Test Phrase 3: This phrase will cause difficulties for all subjects but particularly novices, some of whom will probably give up without completing the phrase. Predictive text will be more difficult to use because of the need to choose the correct word from the dictionary. All words are in the dictionary to avoid the problem of giving every subject a different non-dictionary word to input as the device 'learns' the new word once entered. The full size QWERTY keyboard will perform well for all users with this phrase because the symbols and capital letters are easily identifiable on the keyboard. The miniature QWERTY will also do well though finding some symbols will test some subjects regardless of category. The PDA handwriting will cause problems where case sensitivity is required as the procedure for changing case is not straightforward. The PDA soft keyboard will have similar performance levels to the organisers' QWERTY keyboard but slightly slower because of the smaller size.

Test Phrase 4: This phrase will favour those who normally use the non-predictive method on the phones. This is due to the need to choose the individual letters when using predictive text. The other devices will perform well with this phrase but subject who never abbreviate their messages are likely to spell the full word and so will have a high error rate. The author must make predictions before designing the experiment in order to make the experiment insightful and to ensure the scope of the experiment will achieve its stated goal. It is not necessary for the author's predictions to be right as, if they were, there would be little benefit in conducting the experiment at all.

Conducting the Experiment: Subjects were firstly asked to complete a Pre-experiment Questionnaire to ascertain their experience level. They were then given a brief on each test device and instruction on the use of each. Each subject was given a practice sentence that used every letter of the alphabet at least once and each punctuation mark or symbol required to conduct the experiment. The practice phrase was *The quick brown fox jumped over the lazy dogs.* When the subject was comfortable with the device the test was completed with the same process repeated for each new device.
At the end of each phrase the tester examined the entered text and noted errors in a copybook. The next phrase was then immediately done and the process repeated until all phrases were entered on all devices. At the end of the experiment the subjects were required to complete another questionnaire that asked some simple questions about the test phrases and the devices used. Care was taken to give each subject the device that was deemed to be easier for him or her to use so that the level of complexity increased gradually from one device to the next. When planning the experiment the order of the devices was selected with a preordained assumption of associated difficulty in the opinion of the tester. The spread of test subjects was not arbitrary in nature and consideration was given to selection of subjects so as to give the greatest scope for drawing conclusions and making findings. Factors considered in selection of subjects were age, gender, right or left handed, previous typing/word processing experience, mobile phone text messaging experience, whether Predictive or Non-Predictive text method is normally used by them and type of phone owned. A sample size of 24 subjects was selected as it gave a sufficiently large sample size to encompass the above criteria into the groupings. The groupings were based on experience level of text messaging and selected as follows:

* Novice: These were subjects who rarely or never sent a text message
* Intermediate: These subjects sent between 6 and 15 messages per week
* Expert: Subjects who send more than 15 messages per week.

It was decided that each category should, where possible, contain eight subjects. Half the subjects were female and the other half male. Where possible, each group would have a 50:50 ratio of subjects either side of thirty years of age. The selection of thirty years of age was mostly arbitrary but with consideration for the relatively small sample size and the difficulty of getting subjects in the older age group.

Error Handling: Many papers have been written about predicting text entry speeds but most were concerned with the performance of an expert user. The traditional models were based on Fitts' Law [4] and Keystroke Level Modelling (KLM) [11], which are precise mathematical formulae for finger/hand movement. Dunlop and Crossan [5] expanded the previous studies to model performance of predictive and non-predictive text input devices that allowed for mental preparation time to be taken into account. While their predictions became more accurate when compared to previous studies, the fact remains that there is substitute for large scale empirical studies using real users who make mistakes. There are limitless factors that could be considered when designing an experiment but time and practicality limit us to examining a small number of variables. As this study was a one-man-show, it was decided to limit the examination of errors to a review of the entered text and then comparing it to the given text with time taken being the measured variable. The subjects were briefed on the requirement to correct errors as they occur if seen but not to waste time going back into the text to correct previously unseen errors. This was a major weakness of the experiment as different subjects took varying degrees of care to be accurate rather than speedy. The subject did not know the penalty of an error so that the gamblers among them would not play the system to their advantage by either ignoring errors for speed or speed for errors.
An error in this study is any difference between the given text and the text reproduced by the subject. The Phrases used varied in difficulty and grammatical complexity with punctuation and case sensitivity being critical in some phrases. Any spaces or punctuation marks wrongly included or omitted constituted an error.

Calculating Errors: The equation used here was Soukoreff and McKenzie's Minimum String Distance statistic (MSD) from their study Measuring Errors in Text Entry Tasks: An Application of the Levenshtein String Distance Statistic [10]. MSD is defined as the minimum distance between two strings defined in terms of editing primitives. The primitives are insertion, deletion and substitution. Given two character strings, the idea is to find the smallest set of primitives that applied to one string, produces the other. The number of primitives in the set is the minimum string distance. The actual equation used to calculate the error rate from the MSD is as follows:

$$\text{Error Rate} = \frac{\text{MSD}(A,B)}{\max(|A|,|B|)} \times 100\ \%$$

Where A is the presented text and B is the transcribed text.

This method was favoured, as it does not give undue credit if the subject has less text than was presented while penalising the subject for entering more text than was presented. Below are presented three typical errors encountered with the Error Rate calculated. The three types considered are spelling, insertion and deletion errors. The test phrase used is the simple phrase.

(a) I have never sent a text message before- Presented Text
I have never sent a text message before- Transcribed Text

The length of the transcribed text is 39 characters including spaces. There is one error as underlined. The Error Rate = (1/39) x 100% = 2.56%

(b) I have never sent a text message before-Presented Text
I have never sent text message before-Transcribed Text

The length of the transcribed text is 36 characters including spaces. There are three errors as underlined, i.e. *a* is omitted and two spaces are omitted. The Error Rate = (3/36) x 100% = 8.33%

(c) I have never sent a text message before- Presented Text
I have never sent a*ny* text message before- Transcribed Text

The length of the transcribed text is 41 characters including spaces. There are two errors as underlined, i.e. *ny* is inserted. The Error Rate = (2/41) x 100% = 4.88%.

The method employed below is a variation on the standard MSD error rate method used by Soukoreff and McKenzie [11] in that the length of the given text was related to the device being used so as to account for the extra keystrokes required to get capital letter and symbols. This was done to give a more accurate reflection of the effort involved in different devices. The inclusion of the variable *k* was done by the author in order to calculate errors that involve more than one keystroke per character. However while taking the extra effort of the text input into account for each device, the different effort involved in the making the error was not considered. While this was a failing of the method as outlined in the Anomalous Errors paragraph, it was minor in extent as most mistakes were related to normal

text input. As this method was a variation of the Soukoreff and McKenzie [10, 11] method and therefore could not be directly compared to their results or the results of similar studies, it was therefore deemed necessary by the author to re-calculate the errors while strictly adhering to their method. The only exceptions were that errors in predictive text input were recorded as a single error in cases where the whole word was wrong due to the subject failing to press the choose button to change the word to the required one. In cases where multiple presses of the chose button were required, the relevant number of presses equated to that number of errors. The mathematical equation for the variation on MSD Error Rate is as follows:

$$\text{KSPC Error Rate} = \frac{\text{MSD (A,B)}}{\max k\,(|A|,|B|)+/- k} \times 100\%$$

where, k is the number of keystrokes required to insert punctuation marks and cases.

This is a simple, yet meaningful method of recording errors. Entering extra characters is not penalised as severely as omitting characters as the time factor favours the omission of characters. A single error in a short phrase equates to a much larger error rate than the same error in a longer phrase. This accounts for the sizeable error rates in the data for the simple and abbreviated phrases compared to the error rate for the moderate and complex phrases. This is another draw back of the MSD method for calculating error rate. Overall, the MSD error rate is a useful metric as it allows the subject to enter text in a realistic way correcting errors if they see them but not being obliged to correct all errors or to ignore all errors. The text can then be empirically evaluated while allowing the task to be performed in a natural way.

Anomalous Errors: There are certain types of error that are not weighted fairly by the method chosen by the author. Different devices required more or less keystrokes to produce the given text. Some devices such as the full size QWERTY keyboard require at most two (2) keystrokes to get any character, while the Nokia 6310 phone took up to fifteen keystrokes to get the "+" symbol. Here another weakness of the MSD Error rate becomes obvious as to omit a "+" symbol will give the same error rate as to omit a letter in a word. To account for this error the author would need a Keystroke Per Character (KSPC) value for every character in every phrase for every device and the error specific to that phrase/device would need to be calculated individually. For this reason the author decided to tackle this anomaly by devising the Keystroke Error Rate to compare the results with Minimum String Distance Error Rate. While it would be possible to do this empirically, it would be impractical to do so in the current study. It is an observation worth making and with more manpower and time it would be worth doing to give a more accurate result.

Corrected Errors: Subjects were instructed to correct errors as they are made while at the same time entering the phrase at the best possible speed. Without recording the text entry and error correction to account for time spent fixing errors, it is impossible to differentiate between a fast subject who made and corrected errors and a slower subject who made no mistakes at all. Both would have an error free result with possibly the same time taken by both, or even with the faster typist getting a slower time. To overcome this difficulty would have necessitated the video recording of the subject entering every test phrase into every device. This process was explored initially with the first subject but after thirty minutes recording for one phone it was deemed impractical as the whole experiment for an expert user took over two hours and up to twice that long for a novice. The subject who was recorded expressed unease with the camera and felt it forced errors that would not otherwise have occurred. The time factor for later analysis was also considered, as it would mean spending as much time analyzing the errors as it took to conduct the experiment. As the whole experiment took approximately fifty-two test hours to complete over two months, it was impractical to commit as many hours to recording hours. The weakness of the sum book method of recording errors was that time and keystrokes committed to correcting errors by the subject was not empirically recorded. Some of the uncharacteristically slow times are attributable to this phenomenon. Though this problem was identified early in the experiment it was decided not to tell subjects to correct all errors nor to ignore all errors as it would have lead to confusion for subjects as most people would instinctively correct an error if the saw it. The question of how to continue a word when an error was seen was also considered. Some people would continue the word omitting the intended character that the error related to, while others would retype the intended letter after the error. This would have further complicated the process of error analysis.

RESULTS

Text Entry Speeds: Figure 6-8 showed the performance of all the subjects on selected devices and on each phrase type. Subjects 1-7 are Novices, subjects 8-15 are Intermediates and subjects 16-24 are Experts. Figure 6-8 showed the performance of all subjects on the complex phrase using all devices. The graphs show devices paired with similar devices to aid comparison. This data is raw and takes no account of errors. It is of value, however, as a rough indicator of the speed of input of text in the various devices. The taller the bar in the graph, the longer it takes to input text.

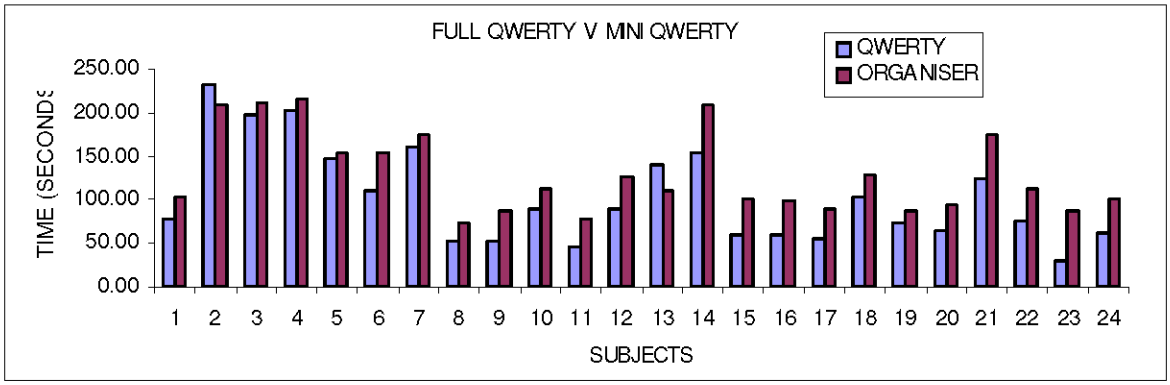

Fig. 6: Complex Phrases on QWERTY Keyboard and Organiser

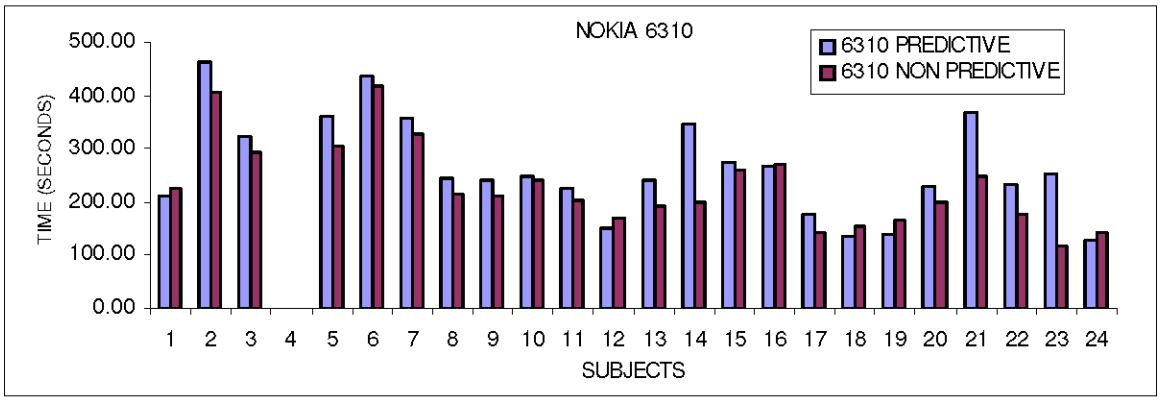

Fig. 7: Complex Phrases on 6310 Phone with/without predictive

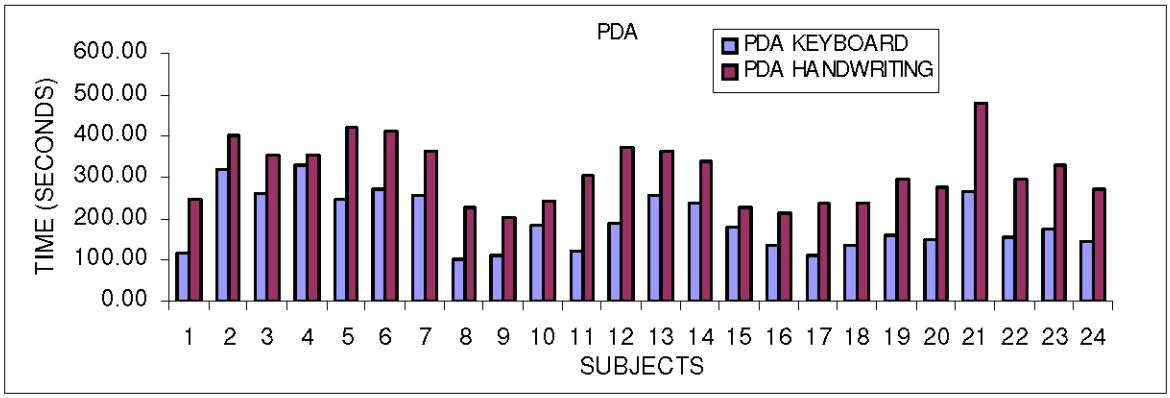

Fig. 8: Complex Phrases on PDA with soft keyboard and hand writing

Figure 6 makes it clear that the full size QWERTY computer keyboard is the fastest means of text input, followed by the mini QWERTY keyboard of the personal organiser and then by the PDA soft QWERTY keyboard. Predictive text entry method is generally quicker than Non-Predictive for both Nokia phones. There were five subjects for whom this was not true, but their questionnaires reveal them to be people who have never used predictive text.

It is noteworthy that the Non-Predictive method is faster than predictive when entering abbreviated phrases. Of the two Nokia phones, the 6310, is on average faster. Of the two text input methods available on the Motorola PDA, the soft QWERTY keyboard with stylus is the fastest by a sizeable margin. Subject 4 was an exception, but he had great difficulty with the small size of the QWERTY keyboard display.

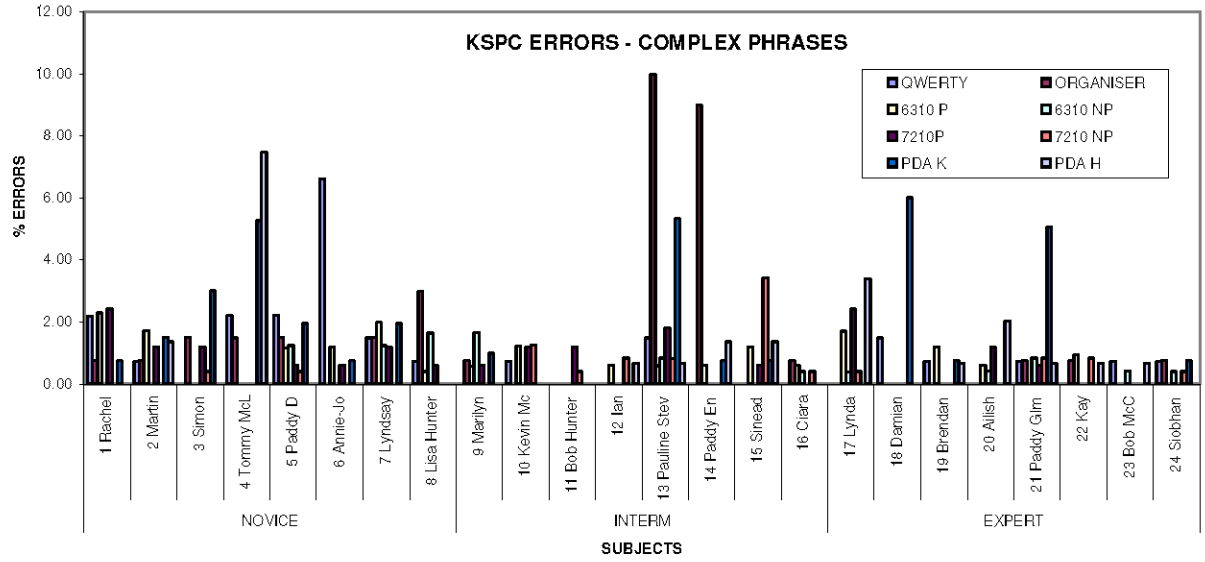

Fig. 9: KSPC Errors for Complex Phrases

Comparing KSPC Errors and MSD Errors: The difference between KSPC and MSD errors was explained earlier. Fig. 9 shows the KSPC errors for all users on the complex phrases. The error range on the Nokia 7210 Non-Predictive for user 15 (Sinead) was found to be 6.78% on MSD, but only 3.57% on KSPC. This is because of the omission of 8 out of 124 characters. The MSD Error equation does not take into account the time factor in making the error.
In this example the omission of eight characters out of a hundred and twenty-four equates to an error twice that of the keystrokes per character error, which takes the KSPC of both the given text and the error into account. The equations to calculate both are:

$$\text{MSD Error\%} = \frac{8}{124-8} \times 100 = 6.78\%$$

$$\text{KSPC Error\%} = \frac{8}{241-17} \times 100 = 3.57\%$$

Anywhere where there are no errors of insertion (inputting too many characters) or omission (leaving out characters), both MSD and KSPC errors are identical. Table 1 shows both MSD and KSPC errors for all subjects, on all phrases and on one device, in this case the PDA in handwriting mode.

Table 1: Average Errors Made by Males and Females on PDA Handwriting

Comparison of Male and Female Subject's Average % Errors				
	Simple	Moderate	Complex	Abbreviated
Male	0.40	0.61	0.89	0.71
Female	0.46	0.67	0.88	0.66

Males made fewer errors on the simple and moderate phrases, but there is no significant difference in males and females for the complex phrase (Table 1). Females were more accurate for the abbreviated phrase. Overall, the difference for males and females is not significant. Table 2 shows the comparison of males and females.

Table 2: Average Errors Made by Young and Old Subjects

Comparison of Old and Young Subject's Average % Errors				
	Simple	Moderate	Complex	Abbreviated
Old	0.42	0.56	0.88	0.72
Young	0.49	0.66	0.93	0.72

Old subjects made fewer errors than young subjects for three of the four phrases and both were equal for the fourth phrase (Table 2). This might have been because the average age of the old subjects is thirty-five years and all were exposed to text input devices at work. Three of the young subjects had little or no exposure to text input devices prior to this experiment and only four of them were graduates.

Input Speed in Words per Minute (WPM): The most accepted empirical measure of text input speed is words per minute. An initial inspection of the results of young versus old shows no significant difference based on age. The order of devices based on speed in words per minute is generally the same as that of the raw data for time taken to input text. As with all the graphs, Novices are slower on average than Intermediates and intermediates are slower than experts. This is not the case for all subjects, with one novice (Rachel) being faster than one expert (Paddy Glm) in most tests. This is a case of 'slow but often' versus 'fast but rarely'. The range of speeds between devices for the males appears

less than it does for females. Females seem to perform significantly better on the full-size QWERTY and mini-QWERTY keyboards than the males do, but the performances on the other devices are broadly similar (Table 3).

Table 3: Average Speeds in Words Per Minute by Sex

Male and Female Subjects' Average Speeds in WPM				
	Simple	Moderate	Complex	Abbreviated
Male	11.57	8.76	7.25	8.45
Female	11.85	8.90	7.37	8.56

Overall, females are on average faster than males, though not significantly. This is possibly due to their greater dexterity with the small devices, or perhaps better hand-eye coordination. The difference in error rates for male and female was almost negligible but with the females making slightly fewer errors. As with the old versus young, the sample size of twelve males and twelve females was too small to make meaningful statistical deductions. To use the statistical method of chi-squared testing, a minimum sample size of thirty subjects is required. Females are generally faster but less accurate than males, in this sample at least. It is possible that females sacrifice accuracy for speed. The sample size would need to be much larger, in the order of thousands, for true statistical analysis to be done on such a broad range of devices. The comparison of young and old doesn't highlight significant differences in terms of speed (Table 4).

Table 4: Average Speed in Words Per Minute by Age

Old and Young Subjects' Average Speeds in WPM				
	Simple	Moderate	Complex	Abbreviated
Old	11.84	9.22	7.37	8.56
Young	11.23	8.54	7.04	8.37

There is a slight bias in terms of the young being marginally slower on average than the old. This seems surprising but is possibly due to the older subjects being exposed to these devices at work or at college and our definition of old as being over thirty years of age is not realistic in real world terms. There is little difference between the different sexes and ages. The difference in error rates for young and old show young subjects to be less accurate than old subjects, but again, only marginally.

Devices: The devices tested are representative of current constrained mobile devices in general use by the population. The findings in chapter 7 graphically illustrate the differences between the devices when used by novices, intermediates and experts on phrases of different complexity. The study revealed some interesting findings about the difference between males and females and young and old, when speed and accuracy are considered for various devices. Table 5 is a summary of the whole sample group where the subjects' performances on all devices are averaged in order to place the devices in order from 1-8 (1 being best and 8 being worst). Table 5 shows how each device is ranked in terms of errors.

The user's favourite was the Nokia 7210 non-predictive. The same phone is ranked last for speed in WPM in the next table, possibly because the difficulty of text input caused users to be more cautious when entering text. The following table shows the ranking of devices in terms of speed in WPM. This is a more accurate method of ranking devices, as subjects' speeds were more constant than their error rates throughout the experiment.

Table 5: Ranking of Devices in Terms of % Errors

Category	Qwerty Keyboard	Organiser MiniQwerty	6310P	6310NP	7210P	7210NP	PDA Qwerty	PDA Handwriting
% Error	0.73	0.78	0.85	0.88	0.88	0.47	1.15	0.99
Rank	2nd	3rd	4th	5th	5th	1st	7th	6th

Table 6: Ranking of Devices in Terms of Speed in WPM

Category	Qwerty Keyboard	Organiser MiniQwerty	6310P	6310NP	7210P	7210NP	PDA Qwerty	PDA Handwriting
Speed WPM	18.71	14.36	7.97	6.94	6.66	6.12	8.05	4.58
Rank	1st	2nd	4th	5th	6th	7th	3rd	8th

Table 7: Ranking of Devices in Terms of Speed in WPM and % Errors

Category	Qwerty Keyboard	Organiser MiniQwerty	6310P	6310NP	7210P	7210NP	PDA Qwerty	PDA Handwriting
Rank % Errors	2nd	3rd	4th	5th	5th	1st	7th	6th
Rank WPM	1st	2nd	4th	5th	6th	7th	3rd	8th
Overall	1st	2nd	3rd	4th	5th	3rd	4th	6th

Table 8: Ranking of Devices in Terms of Subjects Stated Preferences

Device	Qwerty Keyboard	Organiser	Nokia 6310 Mini Qwerty	Nokia 7210	Motorola PDA Qwerty	MotorolaPDA Handwriting
Overall Rank	1st	4th	2nd	6th	5th	3rd

Table 6 shows how the QWERTY keyboard devices fare better in terms of speed, while the phone keypads are similar but with the standard layout of the Nokia 6310 slightly outperforming the non-standard keypad layout of the Nokia 7210. When both sets of results are combined as in Table 7, they reveal the overall ranking of devices.

Table 7 shows the preferences for the devices as stated by the subjects in their Post-experiment Questionnaires. From Table 8, it may be concluded that:

* People prefer the larger keypad/keyboard devices to their smaller equivalents.
* People preferred the standard keypad layout to non-standard layout on phones.
* Opinions on the PDA varied greatly with some loving it while others disliked it.

When speed and accuracy are important, the choice of device would be a small device that has a large QWERTY keyboard. Some of the older subjects (over fifty years old) stated during the experiment that they had difficulty with small keypads and screens. This would be the best compromise, provided the keyboard could fold up into the device. Such devices are under development and use such technology as holographic and flexible LCD screens. Until such technology becomes available and affordable, people can make do with soft, foldout, keyboards that have a docking device for a phone or PDA. While conducting the experiment many test subjects made observations about the devices. These are noted below under the heading of the device about which the observation was made.

Miniature QWERTY Keyboard: Subjects were generally satisfied with this method and had no difficulty using it except perhaps, using the symbols key to access punctuation marks. This was a new departure from the standard QWERTY keyboard.

Nokia 6310 Standard Phone: Almost all subjects found this keypad easy to use both in terms of button size and the size of the print of the letters and numbers on the keys. The main complaint about this phone was the slowness of scrolling through the symbols to get the desired symbol. This was overcome in the Nokia 7210, which uses scroll arrows to move between the lines of symbols. A subject who was familiar with the Nokia 6310 phone showed the tester a shortcut for scrolling through the symbols using the number pad. This option reduced the keystrokes per character (KSPC) for those subjects who used the shortcut. The KSPC was then personalised to the subjects who used the shortcut in order to give a true KSPC Error Rate.

Nokia 7210 Non-Standard Phone: The main observations of the subjects regarding this phone were critical of the small key size and subsequently, almost illegible letter and number size. Most subjects liked the appearance and compactness of the phone, until, that is, they began inputting the test phrases.

Motorola Accompli 008: Soft QWERTY Keyboard: The size of the screen of this device was regarded by all subjects as being too small. The small size limited the size of the QWERTY keyboard which meant good eyesight and dexterity were required to perform well on this device. The location of the 'Cancel' and 'OK' buttons caused confusion also because the natural tendency is to press cancel when an error is made rather than the backspace arrow. If the subject slipped when selecting the symbols key the 'OK' button would be selected. Both these errors blanked the screen, with the former permanently deleting the text already entered. This happened to approximately 20% of the subjects.

Motorola Accompli 008 Handwriting: The small screen size was less of a hindrance when in the handwriting mode, but as with the QWERTY keyboard, the proximity of the 'Cancel' button to the symbols key and the 'OK' button to the space key made fatal errors possible. Left-handed subjects had to draw their 'O's in the opposite direction to normal for the device to recognise the letter. Left-handed subjects accounted for 29% of the test group in this experiment, which is a significant proportion to inconvenience. Generally, those with neat handwriting fared better with this device. Those subjects who hold their pen close to the point occasionally touched the highly sensitive screen, sometimes with fatal consequences (to the test). One subject decided to write the phrases without correcting errors as she went and subsequently corrected all the errors at the end. This proved to be a very slow method because the alternative letter choices were not available once each letter was accepted and the original letter offered may have been offered again. This is an example of poor method affecting speed.

CONCLUSION

There is scope for further examination of the issues raised in this study in terms of usability of devices and speed versus accuracy of text input on these devices. The HCI aspect of design must become more central if

improvements are to be made in this regard. This will require both hardware and software manufacturers to synchronise their efforts to ensure a common goal when designing the next generation of constrained mobile devices. Different levels of application of these principles can then be applied to a particular device in order to customise the device to the target audience. If a device is for informal social communication then more emphasis can be placed on ease and speed of alphanumeric input, whereas if the device is information critical such as a military target identification system, then the accuracy factor becomes crucial.

REFERENCES

1. Buzing, P., 2003. Comparing different keyboard layouts: Aspects of QWERTY, DVORAK and alphabetical keyboards. http://pds.twi.tudelft.nl/~buzing.
2. Butts, L. and Cockburn, 2002. An Evaluation of Mobile Phone Text Input Methods. In: Proc. Third Australasian User Interface Conference (AUIC2002), Melbourne, Australia. Conferences in Research and Practice in Information Technology, Grundy, J. and P. Calder, Eds., ACS, pp: 55-59.
3. James, C. and K. Reischel, 2001. Text input for mobile devices: Comparing model prediction to actual performance. CHI, 31 March-5 April, pp: 366-371.
4. Fitts, P.M., 1954. The information capacity of the human motor system in controlling the amplitude of movement. J. Exptl. Psychol., 47: 381-391.
5. Dunlop, M.D. and A. Crossan, 2000. Predictive text entry method for mobile phones. Personal Technologies, pp: 134-143.
6. Scott MacKenzie, I., H. Kober, D. Smith, T. Jones and E. Skepner, 2001. LetterWise: Prefix-based disambiguation for mobile text input. Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology, pp: 111-113, Nov. 11-14, Orlando, Florida
7. Isokoski, P. and R. Raisamo, 2000. Device independent text input: A rationale and an example. Proceedings of AVI 2000 Conference on Advanced Visual Interfaces, pp: 76-83, ACM, New York.
8. Sirisena, A., 2002. Mobile text entry. pp: 23-33.
9. James, C. and M. Longe, 2000. Bringing text input beyond the desktop. CHI, pp: 111-118, April 1-6, The Hague, Holland.
10. MacKenzie, S. and W.R. Soukoref, 2003. Phrase sets for evaluating text entry techniques. CHI, pp: 754-755, April 5-10, Fort Lauderdale, Florida.
11. Soukoref, W.R. and S.I. MacKenzie, 2001. Measuring errors in text entry tasks: An application of the levenshtein string distance statistic. CHI, pp: 319-320, Seattle, Washington.