Children's Phrase Set for Text Input Method Evaluations

Akiyo Kano
Child Computer Interaction Group
University of Central Lancashire
Preston
United Kingdom
akano@uclan.ac.uk

Janet C Read
Child Computer Interaction Group
University of Central Lancashire
Preston
United Kingdom
jcread@uclan.ac.uk

Alan Dix
Computing Department
Lancaster University
Lancaster
United Kingdom
alan@hcibook.com

ABSTRACT
This paper investigates the suitability of current phrase sets available in HCI for use with children in text entry experiments. It first examines the use of phrase sets within text input method evaluation, and suggests several reasons why the currently available phrase sets may not be suitable for use with children. A new phrase set, containing 500 phrases which have been taken from children's books, is presented. A study that compared the adult focused phrase set with the new children's phrase set is described. This study concludes that the new phrase set is suitable for use with children and, given that results with the two phrase sets were similar, the study adds validity to the existing adult phrase set.

Author Keywords
Phrase set, children, text entry, keyboard, evaluation

ACM Classification Keywords
H.5.5 [Information Interfaces and Presentation]: User Interfaces - Input devices and Strategies, Interaction styles, Evaluation/methodology.

INTRODUCTION
Text input research has being going on for a significant length of time, with its first wave of interest in the 1970s and early 1980s [5]. There is now a wide range of text input methods that cover almost every user need. While there have been many text input method evaluations carried out using adults and university students as participants [14, 6, 3], there are comparatively fewer studies carried out with children. However, nowadays, not only are children required to produce word-processed schoolwork, they also spend considerable leisure time entering text on computers [11].

The work described in this paper is an early study for a much larger project involving the evaluation of text input methods with children. One method that is commonly used in the evaluation of text input methods requires the users to input short phrases. The resulting text is then compared with the phrases that were presented and an error rate is calculated [9, 3]. For the larger project for which this study is a part, the choice of phrase sets to use with children was seen as an early problem; partly as the commonly used phrase sets were designed for adult use, and also as these phrase sets had not been tested for suitability with children.

Methods for Evaluating Text Entry
As new text entry methods are devised, there is a need to evaluate how efficient these methods are. Text entry method evaluation was originally based around typewriters, and researchers used a stopwatch to measure typing speed and counted errors by hand [9].

Modern research has lead to a refinement of methods used to evaluate text input. As alluded to earlier in this paper, in a typical text entry experiment, the participant is shown a short phrase (the intended text), and is then asked to enter the phrase into the device (the transcribed text) while the speed at which it is done and the accuracy with which the text is inputted is measured [1, 8]. The error rate of the inputted text is then calculated by comparing the intended text with the transcribed text. When measuring these variables the text shown to the participant itself becomes an independent variable and careful considerations must be made to ensure that it does not cause any variation in the measurements.

Use of Phrase Sets
There are essentially two choices for text input research. One is to allow the participant to enter "*whatever comes to mind*". This seems to increase the external validity in that the text used is natural to the user. However, as MacKenzie and Soukoreff [9] indicate, it is impossible to measure the accuracy as what the participant *intended* to type is then unknown. Therefore, the second choice, to present the participant with a carefully chosen set of phrases for copying is preferred in most research experiments.

Examples of Currently Available Phrase Sets

Several different strategies have been taken in creating phrase sets, some use newspaper sentences or sentences that are supposed to emulate a conversation [3], others use input phrases that are considered familiar to the user [9]. Phrase sets are generally designed to be moderate in length, easy to remember and representative of the target language. However, as indicated earlier, current phrase sets are designed with adults in mind and their validity with children is untested.

Use of the Phrase Sets with Children

There have been some studies that have used the currently available phrase sets on children [11, 12]. However, due to the adult-orientated content of these phrase sets, the researchers had to pick and choose phrases that were suitable for use with children.. As a result of selecting phrases within the phrase set, the validity of the experiment may be reduced. A need for a child orientated phrase set was identified when this pick and choose method was used repeatedly.

DESIGN OF THE NEW PHRASE SET

Text Entry Phrase Set (TEPS)

The most commonly used phrase set is one created by MacKenzie & Soukoreff [9]. This phrase set, recently named TEPS (Text Entry Phrase Set) by MacKenzie [7] contains 500 phrases with no punctuation symbols and only a few instances of upper case characters which the participants may be instructed to ignore by the researcher during a text entry experiment. The phrase set has been "used in recent studies with good results" [9].

Unsuitability for Use with Children

As stated before, in creating a phrase set, the goal is to use phrases that are moderate in length, easy to remember and representative of the target language. The authors have identified three main categories of problematic phrases in TEPS for use with children.

Unsuitable words for children!!

- `he played a pimp in that movie`
- `make my day you sucker`
- `you are a capitalist pig`

Some words are Americanised

- `my favorite sport is racketball`
- `vanilla flavored ice cream`

Words /terms they may not know

- `the dow jones index has risen`
- `sprawling subdivisions are bad`
- `coalition governments never work`

The New Phrase Set

The new phrase set, Children's Phrase Set (CPSet) is intended to be similar to TEPS, but has been adapted for use with children. The set contains 500 phrases taken from children's books and nursery rhymes [13, 2] and is designed to be suitable for anyone above the age of 6 years old. It contains no capital letters except 'I', no numbers, and no punctuation symbols. It also contains no American or British specific terms. The full phrase set can be obtained from the website: http://www.aks-research.co.uk/CPSet.txt

An analysis of the two phrase sets (TEPS and CPSet) was conducted using AnalysePhrase.java [9] and PHANTIM [4] and the results are shown below in table 1:

	TEPS	CPSet
PHRASE SET		
Num of phrases	500 phrases	500 phrases
Num of words	2713 words	2350 words
Max phrase length	9 words 43 letters	7 words 34 letters
Min phrase length	3 words 16 letters	3 words 12 letters
Average Phrase Length	5.4 words 28.62 letters	4.6 words 22.0 letters
Num of letters	14310 letters	10998 letters
Correlation with English	0.954	0.982
WORDS		
Num of unique words	1164 words	842 words
Max word length	13 letters	11 letters
Min word length	1 letter	1 letter
Average word length	4.46 letters	3.89 letters
Words containing non-letters	0 words	0 words

Table 1: Analysis of the two phrase sets

CPSet contains 363 less words and thus 3312 less letters than the TEPS. Each phrase in the CPSet tends to be on average shorter (by 0.7 words), and thus there are fewer letters in each phrase. There are also less unique words in CPSet (322 words less) but CPSet has a higher correlation with English (using the letter frequencies of Mayzner and Tresselt, [10]), with 0.982 for CPSet and 0.954 for TEPS.

EVALUATING THE NEW PHRASE SET

In designing text input method evaluations, MacKenzie & Soukoreff [9] write:

> *"Among the desirable properties of experimental research are internal validity and external validity. Internal validity is attained if the effects observed are attributed to controlled variables. External validity means the results are generalizable to other subjects and situations"*

This implies that the text entry methods or the devices used become the controlled variable and all other factors should be kept at a constant.

It was therefore important to ensure that, the choice between using TEPS and CPSet did not significantly affect the results of a text input method evaluation i.e. that choosing to use the new phrase set would not cause the participants to create more or less errors.

In order to compare the two phrase sets, a one-day study was carried out involving 40 children from a local junior school. There were 22 boys and 18 girls, aged between 7 and 10 years old. The study was carried out in a quiet room

of a school, using 4 identical black keyboards (PC Line PCL-K350) connected to 4 identical tablet PCs (RM Tablet PC CE0984) on stands (not used as tablets, simply used to create a consistent display). Four children individually carried out the test at a time. Three researchers oversaw the entire study, and there were no video or audio recordings. No names were taken, and results were labeled with only numbers.

Design

In this study the children were asked to copy phrases shown to them on paper, into Notepad™, via a standard QWERTY keyboard. It was decided that each child would type in 10 phrases, as from previous experience, children of this age group tend to lose interest in the task after copying about 10 phrases.

50 phrases were chosen from each phrase set by first randomly choosing a number between 1 and 10, and then selecting every 10^{th} phrase from the phrase set. The children entered 5 phrases from one set, then 5 phrases from the other set. The order of which phrase set was shown first to them was randomized to eliminate any learning effects on their performance. The chosen 100 phrases (50 from TEPS and 50 from CPSet) were each used 4 times in all.

Procedure

Participants were selected by their teachers but guidance was given by the researchers to ensure a representative sample, with respect to age and gender, was used. The children were asked to sit in front of a tablet PC/Keyboard set up and each had the procedure individually explained to them. The children were each given a sheet of paper with the phrases to type in, presented in Arial font, size 20. Children were instructed to copy the phrases that were on the sheet in front of them, into the tablet PC using the keyboard, and told that the trial was not timed, nor was it marked. During the trial, every keystroke was recorded using KGB Keylogger®. Once the child completed the task, he or she left the room and was replaced by another child.

Analysis

Each phrase was tested four times within this study. Afterwards, a count was made of how many INF (Incorrect and Not Fixed), IF (Incorrect and Fixed) and C (Correct) keystrokes was made by each participant for each phrase. Three values, Total Error Rate (TER), Corrected Error Rate (CER) and Not Corrected Error Rate (NCER) [9], were calculated for each of the phrases entered, and then overall for each phrase set. The Total Error Rate was calculated using the formula:

$$Total\ Error\ Rate = \frac{INF + IF}{C + INF + IF}$$

The Corrected Error Rate made for each phrase was also calculated as:

$$Corrected\ Error\ Rate = \frac{IF}{C + INF + IF}$$

The Not Corrected Error Rate made for each phrase was calculated as:

$$Not\ Corrected\ Error\ Rate = \frac{INF}{C + INF + IF}$$

Comparison was then made between the two phrase sets to see if there were any significant differences in the above three values.

Results

Table 2 below shows a summary of the mean TER, CER and NCER for the two phrase sets. Details of each participant's error rates can be found at http://www.aks-research.co.uk/test1.xls

	TEPS		CPSet	
TER	**0.0654**	(sd = 0.0540)	**0.0636**	(sd = 0.0638)
CER	**0.0253**	(sd = 0.0240)	**0.0230**	(sd = 0.0284)
NCER	**0.0400**	(sd = 0.0492)	**0.0340**	(sd = 0.0542)

Table 2: TER, CER and NCER for the two phrase sets

The two phrase sets produced very similar results with regard to error rates. The difference in the TER for the two phrase sets was not significant ($p = 0.829 >> 0.05$, $N = 40$, $t = 0.217$). The differences in CER and NCER for the two phrase sets were also not significant (for CER, $p = 0.321$, $N = 40$, $t = 1.01$ and for NCER, $p = 0.387$, $N = 40$, $t = 0.875$). In each case the difference between TEPS and CPSet is small suggesting performance of the participants in the text input task was not affected by which phrase set the phrases they typed came from. While the between subjects differences were large, a power analysis shows that the minimum discernible effects were: TER 24%, CER 33% and NCER 38%, so that any substantial differences would have been detected.

DISCUSSION

The comparison of TER of the phrase set reveals that the participants made similar numbers of errors and fixed similar numbers of errors for each phrase set. It was also observed that the more errors the child made on phrases from one phrase set, the more errors he or she made on the phrases from the other phrase set (correlation coefficient 0.645, $p<0.05\%$). This suggests that, the two phrase sets can be used interchangeably without lowering the internal validity of a text input method evaluation.

The CER and NCER of the phrase sets confirms this finding; comparison of the CERs show that the children made about the same number of corrections for both phrase sets, and the NCERs show that the children left about the same number of errors uncorrected in their typing. In both

cases strong ($p<0.05\%$) correlations were also found between each child's performances on the two phrase sets.

CONCLUSION

The results show that there were no significant differences in the performance of the participants across the two phrase sets. This indicates that researchers will be able to choose to use the CPSet for text input method evaluations with children, the content of which is more suitable for children, without the choice affecting the results, and thus not lowering the internal validity of their text input method evaluation.

The CPSet is also location independent since it contains no Americanised or British terms, and as it could be used by adults, it offers a common phrase set that can be used in a text method evaluation that involves both children and adults.

An additional finding is that, as the two phrase sets resulted in similar results, the TEPS is further validated as being representative for use.

Future Work

Besides the number of errors that are made by participants, the study has produced a large set of data about the way children perform copying tasks, how they make errors, how they fix these errors, and when they carry out these corrections. Further work is required on this study to understand their behaviors, and also to carry out the same study on older children and adults to see if they behave any differently to the children that were involved in this study.

REFERENCES

1. Butts, L. & Cockburn, A. (2001), An Evaluation of Mobile Phone Text Input Methods, *Third Australasian User Interfaces Conference,* Melbourne, Australia, Conferences in Research and Practise in Information Technology, Vol. 7, pp.55-59.
2. Cooling (Eds.) (1998) "Read Me a Story, Please", Orion Children's Books, UK.
3. James, C. L., & Reischel, K. M., (2001), Text Input for Mobile Devices: Comparing Model Prediction to Actual Performance, *Proceedings of CHI2001*, ACM, pp. 365-371, New York: ACM.
4. Kano, A., (2005), Evaluating Phrase Sets for Use with Text Entry Method Evaluation with Dyslexic Participants, Workshop on Improving and Assessing Pen-Based Input Techniques at British HCI 2005, Edinburgh, England.
5. MacKenzie, I. S., (2002a), Introduction to This Special Issue on Text Entry For Mobile Computing, *Human-Computer Interaction,* Vol.17*,* pp.141-145
6. MacKenzie, I. S. (2002b). Mobile text entry using three keys. Proceedings of the Second Nordic Conference on Human-Computer Interaction – NordiCHI 2002, 27-34. New York: ACM.
7. MacKenzie, I. S., (2006), TEPS, A shortened name for the phrase set, *Personal communication*, June 2006
8. MacKenzie, I. S., & Soukoreff, R. W. (2002), Text entry for mobile computing: Models and methods, theory and practice. *Human-Computer Interaction*, vol. 17, pp. 147-198.
9. MacKenzie, I. S., & Soukoreff, R. W. (2003), Phrase Sets for Evaluating Text Entry Rate Development, *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems – CHI 2003*, pp.754-755. New York: ACM.
10. Mayzner, M. S., and Tresselt, M. E. (1965) Table of singleletter and digram frequency counts for various word-length and letter-position combinations, *Psychonomic Monograph Supplements 1* (1965), 13-32.
11. Read, J. C. (2005), The Usability of Digital Ink Technologies for Children and Teenagers. *HCI2005*, Edinburgh, Scotland, Springer.
12. Read, J. C., MacFarlane, S. J., and Casey, C. (2001), Measuring the Usability of Text Input Methods for Children, HCI2001, Lille, France, Springer-Verlag 2001
13. Schiller, P., & Moore, T., (2004) "Do You Know the Muffin Man? – An Essential Preschool Literacy Resource", Gryphone House, MD.
14. Soukoreff, R. W., & MacKenzie, I. S. (2003). Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. Proceedings of the ACM Conference on Human Factors in Computing Systems – CHI 2003, pp. 113-120. New York: ACM.