Phrase Sets for Evaluating Text Entry Techniques

I. Scott MacKenzie[1,2] and R. William Soukoreff [1]

[1] Dept. of Computer Science
York University
Toronto, Ontario, Canada M3J 1P3
+1 416-736-2100
{smackenzie,will}@acm.org

[2] Unit for Computer-Human Interaction (TAUCHI)
Dept. of Computer & Information Sciences
FIN-33014 University of Tampere
Tampere, Finland
+358 3 215 8566

ABSTRACT

In evaluations of text entry methods, participants enter phrases of text using a technique of interest while performance data are collected. This paper describes and publishes (via the internet) a collection of 500 phrases for such evaluations. Utility programs are also provided to compute statistical properties of the phrase set, or any other phrase set. The merits of using a pre-defined phrase set are described as are methodological considerations, such as attaining results that are generalizable and the possible addition of punctuation and other characters.

TEXT ENTRY EVALUATIONS

Among the desirable properties of experimental research are *internal validity* and *external validity*. Internal validity is attained if the effects observed are attributable to controlled variables. External validity means the results are generalizable to other subjects and situations. Simple as this seems, these attributes are typically at odds with one another. That is, too strictly attending to one tends to compromise the other. This paper pertains to one such point of tension between internal and external validity: the text entered by the participants in evaluations of text entry techniques.

Text entry research typically pits one entry method against another. Thus, *entry method* is the controlled variable, and it is manipulated over two or more levels, for example, *Multitap* vs. *Letterwise* in an experiment comparing text entry techniques for mobile phones [2], or *Qwerty* vs. *Opti* in an experiment comparing soft keyboard layouts [3].

Allowing participants to freely enter "whatever comes to mind" seems desirable, since this mimics typical usage. Such a procedure improves external validity since the results are generalizable. Although of unquestionable merit in gauging the overall usability of a *system* or *implementation*, such methodology also has problems. For one, accuracy is difficult to gauge since there is no source text with which to compare the entered text. Also, the lack of control means performance measurements are coincident with spurious behaviours, such as *pondering* or *secondary tasks*. Thus, sources of variation are present in the dependent variables (e.g., speed or accuracy) that are not attributable to the controlled variable. This compromises internal validly because variations in measurements are, in part, due to other effects.

On balance, the preferred procedure – that used in the majority of research studies – is to present participants with pre-selected phrases of text. Phrases are retrieved randomly from a set and are presented to participants one by one to enter.

Creating a Phrase Set

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.

In a recent paper comparing two soft keyboards, MacKenzie and Zhang [3] used a set of 70 phrases. We recently expanded this set to 500 phrases. A few examples from the set follow:

```
video camera with a zoom lens
have a good weekend
what a monkey sees a monkey will do
that is very unfortunate
the back yard of our house
I can see the rings on Saturn
this is a very good idea
```

We have used the new phrase set with good results in recent studies [1, 5], and wish to share them with the community of text entry researchers via this paper.

The phrases contain no punctuation symbols, and just a few instances of uppercase characters. (Participants may be instructed to ignore case and to enter all characters in lowercase.)

The complete set is available from the authors or directly in http://www.yorku.ca/mack/PhraseSets.zip. Some minor modifications may be necessary to convert spellings to a local dialect (e.g., colour vs. color).

A phrase set should be representative of the target language. The analysis of phrase sets is automated

CHI 2003, April 5-10, 2003, Ft. Lauderdale, Florida, USA.
ACM 1-58113-637-4/03/0004

through a simple Java class called `AnalysePhrases`. Below is an invocation with our 500-phrase set:

```
java AnalysePhrases < phrases2.txt
---------------------------------------
phrases: 500
minimum length: 16
maximum length: 43
average phrase length: 28.61
---------------------------------------
words: 2712
unique words: 1163
minimum length: 1
maximum length: 13
average word length: 4.46
words containing non-letters: 0
---------------------------------------
letters: 14304
correlation with English: 0.9541
---------------------------------------
```

The phases vary from 16 to 43 characters (mean = 28.61). There are 2712 words (1163 unique) varying from 1 to 13 characters (mean = 4.46). The correlation in the last line is with the letter frequencies of Mayzner and Tresselt [4]. The five most frequent letters are as follows:

Letter	Frequency	Probability
e	1523	.1064
t	1080	.0755
o	1005	.0702
a	921	.0644
i	879	.0614

The `AnalysePhrases` program is in the zip file noted above. The file also contains `WordFreq`, a utility to deconstruct the phrases (or any other text file), and output a list of unique words and their frequencies. The output is sorted by frequency or sorted by word. Not surprisingly, 'the' is most frequent word (n = 189). The five most frequent words are as follows:

Word	Frequency	Probability
the	189	.0697
a	108	.0398
is	85	.0313
to	57	.0210
of	54	.0199

The `WordFreq` utility includes several command-line options to control the output. If desired, uppercase characters can be converted to lowercase or words with non-alpha characters can be excluded.

Punctuation and Other Characters

An issue frequently debated in discussions on the evaluation of text entry techniques is whether to include punctuation or other characters in the phrase set. Here we see another point of tension between internal and external validity. The main argument in favour of including such characters is that the evaluation more closely mimics real-life interaction, and, therefore, the results are generalizable. The main argument against is that the entry of non-alpha characters introduces a confounding source of variation in the dependent measures, and, therefore, internal validity is compromised and results are less likely to attain statistical significance. So, should punctuation and other characters be included in the phrase set? It depends. The key issues are explored below.

In designing a controlled experiment, practice dictates that behaviours potentially influencing the dependent variables are controlled, or held constant, except those attributable to the variables under investigation (e.g., entry method). Thus participants' behaviours are constrained to mechanisms that differentiate the interaction techniques. For text entry, the mostly significant point of differentiation is the basic mechanism to enter letters, words, and phrases.

If the techniques under investigation include the same mechanism to enter punctuation and other characters, then it is best to exclude these characters from the interaction, because they do not serve to differentiate the techniques. Instead, they represent an additional and undesirable source of variation.

However, if the techniques under investigation include different mechanisms to enter punctuation and other characters, then including these in the phrases merits serious consideration. If included, they represent an additional source of variation, and therefore reduce the likelihood of attaining statistically significant results. One possible remedy is to include "character set" as an additional factor in the design of the experiment, with "alpha-only" and "alpha-plus-punctuation" as the levels.

Acknowledgement

This research is sponsored by the Natural Sciences and Engineering Research Council of Canada and the Academy of Finland (project 53796).

References

1. MacKenzie, I. S. Mobile text entry using three keys, *Proc. NordiCHI 2002*. New York: ACM, 2002, 27-34.

2. MacKenzie, I. S., Kober, H., Smith, D., Jones, T., and Skepner, E. LetterWise: Prefix-based disambiguation for mobile text input, *Proc. UIST 2001*. New York: ACM, 2001, 111-120.

3. MacKenzie, I. S., and Zhang, S. X. The design and evaluation of a high-performance soft keyboard, *Proc. CHI '99*. New York: ACM, 1999, 25-31.

4. Mayzner, M. S., and Tresselt, M. E. Table of single-letter and digram frequency counts for various word-length and letter-position combinations, *Psychonomic Monograph Supplements 1* (1965), 13-32.

5. Soukoreff, R. W., and MacKenzie, I. S. Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric, *Proc. CHI 2003*. New York: ACM, (to appear).