Performance Comparisons of Phrase Sets and Presentation Styles for Text Entry Evaluations

Per Ola Kristensson
School of Computer Science
University of St Andrews
St Andrews, Fife, United Kingdom
pok@st-andrews.ac.uk

Keith Vertanen
Department of Computer Science
Montana Tech of the University of Montana
Butte, Montana, USA
kvertanen@mtech.edu

ABSTRACT

We empirically compare five different publicly-available phrase sets in two large-scale ($N = 225$ and $N = 150$) crowdsourced text entry experiments. We also investigate the impact of asking participants to memorize phrases before writing them versus allowing participants to see the phrase during text entry. We find that asking participants to memorize phrases increases entry rates at the cost of slightly increased error rates. This holds for both a familiar and for an unfamiliar text entry method. We find statistically significant differences between some of the phrase sets in terms of both entry and error rates. Based on our data, we arrive at a set of recommendations for choosing suitable phrase sets for text entry evaluations.

Author Keywords

Text entry, phrase sets, crowdsourcing, keyboards

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation: User Interfaces—Input devices and strategies

General Terms

Experimentation, Standardization

INTRODUCTION

Text entry methods are typically evaluated using a transcription task (sometimes called a copy-task). In this task participants are instructed to write stimulus phrases as quickly and as accurately as possible. Until recently there was no widely agreed upon standard set of phrases for use in text entry evaluations. As a consequence, different researchers tended to choose different text sources. For example, in a study of speech interfaces "text was drawn from an old western novel" [3]. In a study of device-independent text entry methods "phrases were extracted from the fortune cookie database delivered with Red Hat Linux 5.2" [1]. In a study on optimized keyboards "test sentences were randomly selected from news" [8]. MacKenzie and Soukoreff [4] helped improve the situation by contributing a standard set of 500 phrases.

IUI'12, February 14–17, 2012, Lisbon, Portugal.

However, does the choice of phrase set actually matter? This question has been discussed a number of times in the recent text entry literature. MacKenzie and Soukoreff [4] raise the issue of external and internal validity of phrase sets and argue that phrase sets should be standardized. Paek and Hsu [5] follow up on MacKenzie and Soukoreff's argument of the importance of using representative phrase sets and propose a method for sampling representative phrase sets from large bodies of text. Recently, we released a new phrase set based on genuine mobile emails that have been validated for memorability [7]. We argue in [7] that this set's real-world data and demonstrated memorability should increase both internal and external validity in text entry evaluations.

We believe the choice of phrase set *may* matter for the following reasons:

Reproducibility Text entry experiments that use non-standard phrase sets often end up being unreproducible since text sources used for stimuli cannot be located (e.g. text from an unnamed western novel).

Study heterogeneity Empirical measurements from different text entry studies are sometimes compared in systematic reviews. If researchers use different phrase sets they introduce a potential confound making meta-analyses more difficult.

Internal validity In an experiment it is critical that manipulation of the independent variable (the text entry method) is the only major source of variation in the measured dependent variables (typically entry and error rates). Factors that may pose threats to internal validity include phrases that are hard to remember and phrases that contain arcane punctuation or auxiliary symbols.

External validity The relative performance of the text entry methods should ideally also hold when users are composing their own text outside the laboratory. However, a phrase set that is not representative of the text that end-users are likely to write may potentially generate misleading results.

Researchers have so far primarily concentrated on the trade-off between internal and external validity (e.g. [4, 7]), and on how well the phrases model the text end-users are likely to compose (e.g. [2, 4, 5, 7]). In this paper we instead focus on the actual empirical performance implications of phrase sets and presentation styles. We carried out two experiments. Both of these experiments used an identical study design with one between-subjects independent variable (phrase set) and one within-subjects independent variable (presentation style).

Phrase set	Examples
MACKENZIE	fish are jumping great disturbance in the force I took the rover from the shop
MOBILEEMAIL	Is she done yet? How are you? We are all fragile.
AAC	let me see that that's the way its always been where is the restroom
NGRAM	Ref: Tie Line Access because it did not members, you may want
SMS	Like a personal sized or what How do you plan to manage that ya i know liao got loophole

Table 1. Example phrases from the five different phrase sets.

Phrase Sets

We investigated five phrase sets which were selected based on the criterion that the set, or the source for the set, must be publicly available (this ruled out the children's phrase set [2] and the Twitter and Facebook n-gram phrase sets [5]). Table 1 shows some example phrases from each of the phrase sets.

MACKENZIE This phrase set was released by MacKenzie and Soukoreff [4] in 2003. It contains phrases with no punctuation and limited capitalization. 500 phrases, 2713 words, 14305 characters.

MOBILEEMAIL This phrase set was released by Vertanen and Kristensson [7] in 2011. This phrase set contains phrases drawn from genuine mobile emails in the Enron email corpus. The phrase set release contains several subsets. We used the subsets `mem1-mem5` which had been verified to be memorable. 200 phrases, 1073 words, 5253 characters.

AAC A publicly available collection of short conversational messages designed by Augmented and Alternative Communication (AAC) specialists.[1] 952 phrases, 3843 words, 17169 characters.

NGRAM This phrase set was released by Peak and Hsu [5] in 2011. It consists of 4-grams sampled from the Enron email corpus. 500 phrases, 1998 words, 11282 characters.

SMS We created this phrase set from the National University of Singapore SMS corpus.[2] We used their publicly released SMS corpus dated June 20th, 2011.[3] The dataset contained noise and many of the original SMS messages were unintelligible. We filtered the data to include only messages that satisfied two criteria. First, every word in a message had to exist in a large word list. We created the word list by merging Wiktionary, Webster's dictionary provided by Project Gutenberg, the CMU pronouncing dictionary and GNU aspell. Second, a message must consist of 21–74 characters and 4–14 words. These thresholds mirror the minimum and maximum number of characters and words occurring in the MOBILEEMAIL phrase set. 769 phrases, 5442 words, 25261 characters.

[1] `http://aac.unl.edu/vocabulary.html`
[2] `http://wing.comp.nus.edu.sg/SMSCorpus/`
[3] `smsCorpus_en_2011.06.20.xml`

Phrase set	MEMORIZE	TRANSCRIBE
MACKENZIE	69.9 (22.0)	68.6 (22.1)
MOBILEEMAIL	70.1 (17.3)	66.7 (16.9)
AAC	88.0 (25.7)	83.8 (28.4)
NGRAM	67.1 (17.6)	64.6 (20.5)
SMS	62.7 (18.0)	61.0 (16.7)

Table 2. Mean entry rates (wpm) in the familiar text entry method experiment (standard deviations in parentheses).

Presentation Styles

Normally in a transcription-based text entry evaluation, participants can see the stimulus phrase while entering the text. While this avoids an explicit memorization step, it also somewhat lowers participants' entry rates [6]. We therefore decided to also investigate two presentation styles:

MEMORIZE Participants were shown a phrase and asked to try to remember it. After clicking a button, the phrase disappeared and participants typed the phrase from memory. Participants were told to type as much as they could remember.

TRANSCRIBE Participants were presented with a phrase but the phrase remained visible throughout the text entry task.

EXPERIMENT 1: FAMILIAR TEXT ENTRY METHOD

We recruited 225 participants via Amazon Mechanical Turk. We restricted the task to workers resident in the United States. Participants were randomly assigned to one of the five phrase sets. We ensured that each phrase set had an equal number of 45 participants. Each participant was given a random subset of phrases selected from the participant's assigned phrase set.

Participants were presented with a phrase and asked to type it as quickly and as accurately as possible. The session was divided into two parts. The first part consisted of 25 phrases that were presented using one presentation style, either MEMORIZE or TRANSCRIBE (described in the previous section). The second part consisted of 25 phrases that were presented using the other presentation style. The order of the presentation styles was balanced across the participants.

Results

We removed workers who provided obvious garbage responses. After this removal, we had 50 phrases from 225 participants, 11250 phrases in total.

Entry Rates

Entry rates were measured in words-per-minute (wpm), with a word defined as five consecutive characters. Timing was measured as the interval between the first key press and the last key press. Table 2 provides a summary of the entry rates. Participants wrote faster in MEMORIZE (mean = 71.6 wpm, sd = 22.0 wpm) than in TRANSCRIBE (mean = 68.9 wpm, sd = 22.6 wpm). Repeated measures analysis of variance using an initial (unadjusted) significance level of $\alpha = 0.05$ revealed a

Phrase set	MEMORIZE	TRANSCRIBE
MACKENZIE	1.7% (3.4%)	0.2% (0.5%)
MOBILEEMAIL	2.1% (5.9%)	0.2% (0.5%)
AAC	1.1% (2.4%)	0.4% (0.7%)
NGRAM	3.3% (2.3%)	0.6% (0.6%)
SMS	5.5% (6.0%)	0.5% (0.8%)

Table 3. Mean error rates (CER) in the familiar text entry method experiment (standard deviations in parentheses).

Phrase set	MEMORIZE	TRANSCRIBE
MACKENZIE	10.0 (2.3)	9.6 (2.4)
MOBILEEMAIL	9.2 (2.2)	8.7 (2.2)
AAC	11.5 (3.4)	10.8 (2.9)
NGRAM	9.8 (2.2)	9.4 (2.0)
SMS	9.6 (2.6)	10.0 (2.4)

Table 4. Mean entry rates (wpm) in the unfamiliar text entry method experiment (standard deviations in parentheses).

significant difference for presentation style ($F_{1,220} = 29.009$, $\eta_p^2 = 0.116$, $p < 0.001$).

We also found significant differences between the phrase sets ($F_{4,220} = 8.982$, $\eta_p^2 = 0.140$, $p < 0.001$). Tukey HSD *post hoc* tests showed that the AAC set resulted in significantly faster entry rates than all other sets. No other differences in entry rates between the sets were significant. There was no significant interaction between presentation style and phrase set ($F_{4,220} = 1.194$, $\eta_p^2 = 0.021$, $p = 0.314$).

Error Rates

Error rates were measured as character error rate (CER). CER is the minimum edit distance between the participant's written response text and the stimulus phrase, divided by the number of characters in the stimulus phrase. Table 3 provides a summary of the error rates. The error rate was higher in MEMORIZE (mean = 2.7%, sd = 4.5%) than in TRANSCRIBE (mean = 0.4%, sd = 0.6%). A repeated measures analysis of variance using an initial (unadjusted) significance level of $\alpha = 0.05$ showed this was a significant difference ($F_{1,220} = 70.877$, $\eta_p^2 = 0.244$, $p < 0.001$).

We also found significant differences between the phrase sets ($F_{4,220} = 7.569$, $\eta_p^2 = 0.121$, $p < 0.001$). We found a significant interaction between presentation style and phrase set ($F_{4,220} = 6.629$, $\eta_p^2 = 0.108$, $p < 0.001$). We therefore split the dataset by presentation style and performed two separate between-subject ANOVAs with appropriately adjusted significance levels to guard against the risk of over-testing the data. For the MEMORIZE presentation style there was a significant difference between the phrase sets ($F_{4,220} = 7.215$, $\eta_p^2 = 0.116$, $p < 0.0001$). Tukey HSD *post hoc* tests showed that the SMS phrase set resulted in significantly more errors than the MOBILEEMAIL, MACKENZIE and AAC phrase sets. No other differences in error rate between the phrase sets were statistically significant. For the TRANSCRIBE presentation style the differences between the phrase sets were not significant after adjustment of the significance level to guard against over-testing ($F_{4,220} = 3.397$, $\eta_p^2 = 0.058$, $p = 0.01$).

Total Task Time

We also investigated the total task time between the MEMORIZE presentation style and the TRANSCRIBE style. In MEMORIZE, the total task time was measured from when participants were exposed to the phrase to be memorized to the point when they had completed writing the phrase. In TRANSCRIBE, the total task time was measured from when they were exposed to the phrase to the point when they had completed writing the phrase. We found that MEMORIZE resulted in longer total task times (mean = 10.0 s, sd = 5.3 s) than TRANSCRIBE (mean = 7.0 s, sd = 3.3 s). The difference was statistically significant ($F_{1,220} = 199.0$, $\eta_p^2 = 0.474$, $p < 0.001$). In other words, participants had a higher entry rate when they memorized a phrase beforehand. However, they also spent more time performing each task.

Worker Demographics

Before the text entry task, we asked workers a number of questions about themselves. 71% of our participants were female with an average age of 33. The vast majority were native English speakers (93% native speakers, 5% advanced, and 2% moderate/beginners). 56% of workers reported using a laptop, 43% a desktop, and 1% a mobile/tablet device.

EXPERIMENT 2: UNFAMILIAR TEXT ENTRY METHOD

The previous experiment investigated performance differences when participants used their familiar full-sized keyboards. However, in practice most text entry evaluations are conducted using text entry methods that are unfamiliar to participants. We investigated if an unfamiliar text entry method would change the results. In experiment two, participants entered text by pushing buttons on the ATOMIK on-screen keyboard [8]. It has been established that optimized on-screen keyboards have a long learning curve [8]. Thus, even though we could not control the exact pointing device used by our workers, their overall task time in a short experiment would be dominated by the visual search task [8].

The method was identical to the first experiment except for the following. We recruited 150 participants via Mechanical Turk. Each phrase set had an equal number of 30 participants. Participants wrote ten phrases in each presentation style.

Results

As in experiment 1, we removed workers who provided obvious garbage responses. After removal, we had 20 phrases from 150 participants, 3000 phrases in total.

Entry Rates

Entry rates were measured as in experiment one. Table 4 provides a summary of the entry rates. Entry rate was slightly faster for MEMORIZE (mean = 10.0 wpm, sd = 2.7 wpm) compared to TRANSCRIBE (mean = 9.7 wpm, sd = 2.5 wpm). Repeated measures analysis of variance showed that this difference was statistically significant ($F_{1,145} = 5.426$, $\eta_p^2 = 0.036$, $p = 0.021$). As shown in Table 4, faster entry in MEMORIZE held for all phrase sets except for the SMS set.

We found that there were statistically significant differences between the phrase sets ($F_{4,145} = 3.6$, $\eta_p^2 = 0.09$, $p = 0.008$). Tukey HSD *post hoc* tests showed that the AAC

Phrase set	MEMORIZE	TRANSCRIBE
MACKENZIE	4.3% (5.8%)	1.5% (2.5%)
MOBILEEMAIL	5.1% (6.5%)	1.7% (2.5%)
AAC	4.9% (8.0%)	1.6% (2.8%)
NGRAM	4.9% (5.0%)	1.1% (1.9%)
SMS	10.8% (10.9%)	1.7% (3.1%)

Table 5. Mean error rates (CER) in the unfamiliar text entry method experiment (standard deviations in parentheses).

phrase set resulted in significantly faster entry rates than MOBILEEMAIL. No other differences in entry rates between the phrase sets were statistically significant. There was no significant interaction between presentation style and phrase set ($F_{4,145} = 1.659$, $\eta_p^2 = 0.044$, $p = 0.163$).

Error Rates

Error rates were measured as in experiment one. Table 5 provides a summary of the error rates. The error rate was higher for MEMORIZE (mean = 6.0%, sd = 7.8%) compared to TRANSCRIBE (mean = 1.5%, sd = 2.7%). Repeated measures analysis of variance showed this difference was statistically significant ($F_{1,145} = 51.759$, $\eta_p^2 = 0.263$, $p < 0.001$). In MEMORIZE, the error rate was markedly higher for the SMS phrase set compared to the other phrase sets.

There was a significant interaction between presentation style and phrase set ($F_{4,145} = 3.493$, $\eta_p^2 = 0.088$, $p = 0.009$). We therefore split the dataset by presentation style and performed two separate between-subject ANOVAs with appropriately adjusted significance levels to guard against the risk of over-testing the data. For the MEMORIZE presentation style there was a significant difference between the phrase sets ($F_{4,145} = 3.821$, $\eta_p^2 = 0.095$, $p = 0.006$). Tukey HSD *post hoc* tests showed that the SMS phrase set resulted in significantly more errors than all other phrase sets. No other differences in error rate between the phrase sets were statistically significant. For the TRANSCRIBE presentation style the differences between the phrase sets were not significant ($F_{4,145} = 0.209$, $\eta_p^2 = 0.006$, $p = 0.933$).

Total Task Time

Total task times were measured as in experiment one. Total task times were longer for MEMORIZE (mean = 47.1 s, sd = 53.1 s) than for TRANSCRIBE (mean = 44.5 s, sd = 19.0 s). However, the difference was not statistically significant ($F_{1,145} = 0.339$, $\eta_p^2 = 0.002$, $p = 0.561$).

DISCUSSION AND RECOMMENDATIONS

Our MEMORIZE condition provides empirical evidence that the MACKENZIE, AAC and NGRAM phrase sets are indeed memorable by users. Previously we only knew this for the MOBILEEMAIL phrase set. Further, we found that the MEMORIZE presentation style resulted in consistently higher entry rates than the traditional TRANSCRIBE presentation style. However, this came at the cost of slightly higher error rates and somewhat longer task times.

The phrase set by MacKenzie and Soukoreff [4] has been popular for text entry evaluations since its release in 2003. Given that two new phrase sets have recently appeared that may be more suitable in certain situations, such as mobile text [7] or special use-cases [5], it is reassuring that the MacKenzie and Soukoreff [4] phrase set provides similar performance.

We caution against using the SMS phrase set since it was significantly more error prone than the other sets. We believe this was due to the strange language, abbreviations, and sentence fragments in this set. This is supported by comments made by workers at the end of the experiment (e.g. "The ones spelled incorrectly were the hardest", "Bad grammar! Nonsensical!"). In addition, we caution against using the AAC phrase set since it resulted in significantly higher entry rates than the other sets. We believe this resulted from the set consisting almost entirely of simple, short, and familiar phrases which avoided proper names, unusual vocabulary, and difficult grammar. This probably led to faster typing due to factors such as motor memory and ease of memorization.

Based on our data we recommend text entry researchers first consider using either the MACKENZIE or the MOBILEEMAIL phrase sets. Both of these sets perform similarly in terms of entry and error rates for both the familiar and the unfamiliar text entry methods we tested. For mobile text entry methods it may be worth using the MOBILEEMAIL phrase set since it has higher external validity given that it is based on genuine mobile emails. We hope researchers will consider these recommendations when evaluating new text entry methods.

ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council (grant number EP/H027408/1) and the Scottish Informatics and Computer Science Alliance.

REFERENCES

1. Isokoski, P., and Raisamo, R. Device independent text input: a rationale and an example. In *Proc. AVI 2000*, ACM Press (2000), 76–83.
2. Kano, A., Read, J. C., and Dix, A. Children's phrase set for text input method evaluations. In *Proc. NordiCHI 2006*, ACM Press (2006), 449–452.
3. Karat, C.-M., Halverson, C., Horn, D., and Karat, J. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proc. CHI 1999*, ACM Press (1999), 568–575.
4. MacKenzie, I. S., and Soukoreff, R. W. Phrase sets for evaluating text entry techniques. In *Ext. Abstracts CHI 2003*, ACM Press (2003), 754–755.
5. Paek, T., and Hsu, B.-J. P. Sampling representative phrase sets for text entry experiments: a procedure and public resource. In *Proc. CHI 2011*, ACM Press (2011), 2477–2480.
6. Soukoreff, R. W., and MacKenzie, I. S. Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. In *Proc. CHI 2003*, ACM Press (2003), 113–120.
7. Vertanen, K., and Kristensson, P. O. A versatile dataset for text entry evaluations based on genuine mobile emails. In *Proc. MobileHCI 2011*, ACM Press (2011), 295–298.
8. Zhai, S., Sue, A., and Accot, J. Movement model, hits distribution and learning in virtual keyboarding. In *Proc. CHI 2002*, ACM Press (2002), 17–24.