A Versatile Dataset for Text Entry Evaluations Based on Genuine Mobile Emails

Keith Vertanen
Princeton University
Department of Computer Science
vertanen@princeton.edu

Per Ola Kristensson
University of Cambridge
Computer Laboratory
pok21@cam.ac.uk

ABSTRACT
Mobile text entry methods are typically evaluated by having study participants copy phrases. However, currently there is no available phrase set that has been composed by mobile users. Instead researchers have resorted to using invented phrases that probably suffer from low external validity. Further, there is no available phrase set whose phrases have been verified to be memorable. In this paper we present a collection of mobile email sentences written by actual users on actual mobile devices. We obtained our sentences from emails written by Enron employees on their BlackBerry mobile devices. We provide empirical data on how easy the sentences were to remember and how quickly and accurately users could type these sentences on a full-sized keyboard. Using this empirical data, we construct a series of phrase sets we suggest for use in text entry evaluations.

Author Keywords
Text entry evaluation, mobile email, phrase set

ACM Classification Keywords
H5.2. Evaluation/methodology

General Terms
Measurement, Experimentation, Human Factors

INTRODUCTION
Mobile text entry methods are nearly always evaluated by measuring entry and error rates in a text-copy task. Participants are given a series of texts that they must enter "quickly and accurately". A popular test set used in text entry experiments is the MacKenzie and Soukoreff phrase set [4]. This set consists of short memorable phrases such as "the capital of our nation", "a fox is a very smart animal" and "great disturbance in the force". However, phrases in this set were not written by mobile device users and are most likely not very representative of actual mobile messages. Further, it is unknown whether these phrases are in fact easy for participants to remember.

In this paper we describe a versatile, reasonably large, and high quality dataset that contains actual messages written by users on mobile devices. It contains phrases and sentences with varying text lengths, mixed case, symbols and numbers. The dataset also contains extensive and previously unavailable metadata, such as message category (personal, business or Enron-specific), empirical data on memorability, and empirical data on entry and error rates obtained on full-sized keyboards. Further, we provide a script that enables researchers to easily create specific test sets, such as those containing only easy to remember business sentences.

We created the dataset by identifying mobile messages from the publically available Enron email corpus [3]. We used messages written by users on BlackBerry mobile devices. Identifying the mobile messages was possible because by default the BlackBerry appends a standard email signature. We found that 44 of the 150 Enron employees in the corpus had created messages on a BlackBerry and had not disabled the default signature. From these users' messages, we obtained a total of 2239 sentences and sentence fragments (such as opening greetings). Henceforth we use the term sentence to mean both complete sentences and sentence fragments. We manually reviewed the sentences to correct misspellings, to remove repeated messages, and to discard incomprehensible text. We each reviewed half the messages and proofread the work of the other. Table 1 shows a selection of sentences from the set. We have made the dataset publicly available [1].

Thanks, I will look at it tonight.
I'm at the doctor's office this AM, but will be in the office later.
Interesting, are you around for a late lunch?
Davis had not yet updated the model for this.
I'm hoping I can bring Alina and make a play date for Chad.
Thanks DS
Are you going to join us for lunch?
Thanks for the surprise.
How about 9 in my office on 3825?

Table 1: Example sentences from the dataset.

Subset	Number	Description
-	2239	Full test set
simple	2109	Limited punctuation to ?!,.'
nospell	1877	Removed possible acronyms
body	1446	4 or more words, must end in ?!.
nonum	1347	Removed number characters

Table 2: Subsets that progressively filtered out sentences in the dataset.

MobileHCI 2011, Aug 30–Sept 2, 2011, Stockholm, Sweden.

Name	Sentences	Words	Words / sentence	Letters / word	OOV	1st person	Question
MOBILE	**2239**	**20.5K**	**9.1**	**4.1**	**1.4%**	**39.6%**	**15.1%**
NEWS	60.4M	1323M	21.9	4.9	2.2%	8.8%	1.3%
ENRON	1.1M	19M	16.7	4.7	1.9%	27.2%	6.8%
MACKENZIE	500	2.7K	5.4	4.5	0.6%	12.2%	n/a

Table 3: Statistics about our dataset (top row, boldfaced) and three other datasets.

THE DATASET

We defined four subsets of our full dataset that progressively filtered out sentences. These subsets are intended to provide text appropriate to the different capabilities of the text entry interface being tested. For example, some interfaces may lack the ability to enter symbols or numbers. Table 2 summarizes these subsets. To aid evaluations of speech interfaces, we also include a version of each subset with verbalized numbers, punctuation and spelling (e.g. "i heard it was at five ?question-mark") and another version that drops all punctuation (e.g. "i heard it was at five").

We compared our dataset (MOBILE) against a number of other datasets:

- NEWS – News text from the CSR and Gigaword corpora.
- ENRON – Enron corpus excluding emails used in MOBILE.
- MACKENZIE – The MacKenzie and Soukoreff phrase set (phrases2.txt) containing 500 short phrases [4].

The reported out-of-vocabulary (OOV) rate is with respect to the most frequent 64K words in the WSJ0 corpus. Sentences were counted as being in the first person if they contained a first person personal pronoun (me, my, mine, myself, I, I'm, I've, I'd, I'll, our, ours, ourself, ourselves, we're, we'd, we'll, we've). Sentences ending in a question mark were considered questions. Punctuation was removed before we calculated the statistics.

As shown in Table 3, the MOBILE sentences were markedly different from the other text genres. This included even the other email messages in the Enron corpus. MOBILE sentences were very short with an average length of nine words. Compared to other datasets, the MOBILE set had a much higher percent of sentences in the first person (40%) and sentences that were questions (15%).

We placed sentences into one of three categories: business, personal, or Enron specific. We found 51% of sentences were personal, 30% personal, and 19% were Enron specific. The main advantage of this categorization is to allow removal of Enron specific sentences. While some sentences were clearly personal or business, many could plausibly be in either category (e.g. "So the language does have value").

SENTENCE MEMORABILITY

Each sentence in the dataset is supplemented with information about how easy it might be to memorize. In general, copy-tasks should prefer memorable stimuli. However, until now, no datasets have had empirical information regarding the memorability of stimuli. For text entry methods requiring constant visual attention (such as eye typing), memorability is particularly important since it can be difficult for participants to refer to a reference text.

We obtained an empirical estimate of memorability by running an experiment on the crowdsourcing site Amazon Mechanical Turk. In our experiment workers read a sentence and then typed the sentence from memory after it was removed. Our experiment asked workers to first rate their English ability and specify their country. They were then shown a sentence from our test set (Figure 1 left). Workers were instructed to try to remember the sentence since it would disappear once they pressed the continue button. After pressing the continue button, the worker was prompted to type in the previously displayed sentence as accurately as possible from memory (Figure 1 middle).

Workers were told not to write anything down and were prevented from pasting text into the result text area. Each human intelligence task (HIT) consisted of a set of 20 sentence tasks and workers were paid $0.10 for the HIT. Each HIT was completed by ten different workers. This provided multiple measurements for each sentence in our test set. For simplicity, we performed the experiment on all our sentences including some that were blatantly too long for successful memorization. Our HIT was instrumented to record the times of all keyboard and button actions.

Workers were shown the original text which often contained mixed case, punctuation, and sometimes

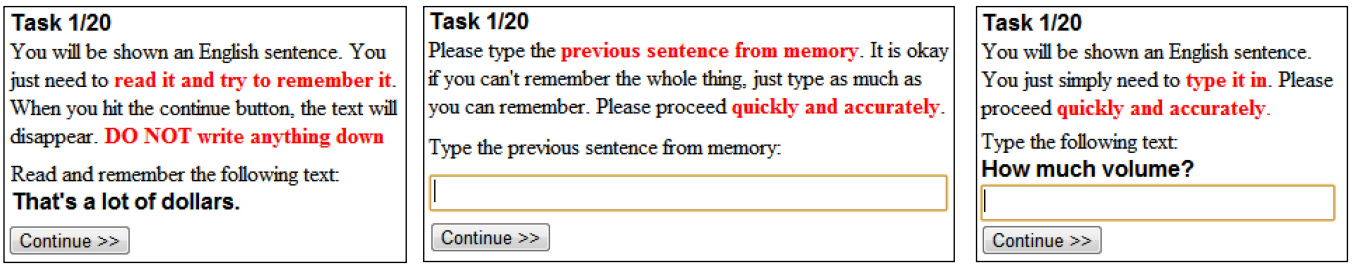

Figure 1: The instructions given for the memorization experiment (left, middle) and the entry rate experiment (right).

numbers. We computed how often they typed the target sentence exactly correct. Since we were curious how often workers made mistakes with regard to capitalization and punctuation, we also computed the percent correct ignoring case and ignoring both case and punctuation.

A total of 22,390 sentence memorization tasks were completed by 386 unique workers. In 51% of the tasks, workers typed the exact target sentence. As shown in Figure 2, the percent correct decreased as sentences got longer. Differences in case constituted about 5% of the errors and differences in punctuation constituted another 15% of the errors. For short sentences of 1–5 words, ignoring case and punctuation errors, workers typed on average 89% of the sentences correctly. Such short sentences would be prime candidates for use in evaluations requiring quick memorization of target texts. As shown in Figure 3, the time spent reading and attempting to memorize a sentence increased as sentences became longer.

We observed that performance depended on the worker's country. Most tasks were completed either by workers from India (64%) or from the United States (26%). Indian workers got 45% of sentences completely correct while US workers got 62% correct. We suspect this is due to differing English language abilities. Indeed workers self-reporting to be native English speakers got 59% correct while those reporting to be beginners at English got 44% correct.

SENTENCE ENTRY AND ERROR RATE

We also supplemented each sentence with empirical entry and error rate estimates. These rates were obtained by conducting a second experiment in which workers performed a text-copy task using full-sized keyboards. This experiment was similar to our first except workers could see the target sentence while typing (see Figure 1 right).

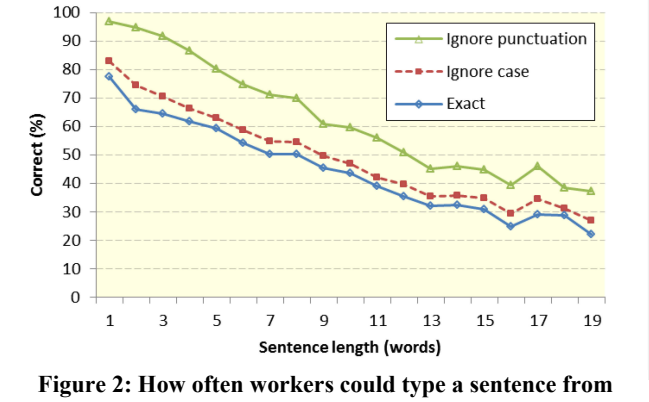

Figure 2: How often workers could type a sentence from memory completely correctly for different length sentences.

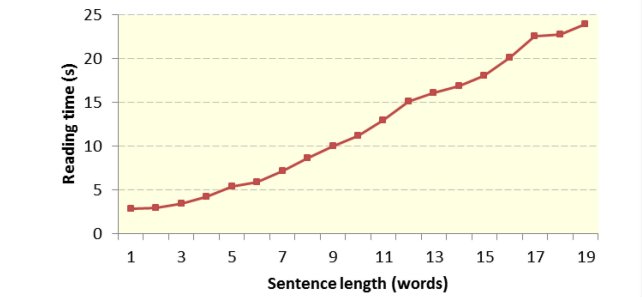

Figure 3: Time spent reading in the memorization experiment.

We computed the entry rate in words per minute (wpm). The number of words was calculated using the standard convention of dividing the number of characters including spaces by five. Timing was done from the first key press in the text result box until the last key press. Overall the entry rate was 50 wpm. As shown in Figure 4, entry rate became faster as the sentence length increased from one word to six words. Thereafter, entry rates remained fairly constant.

We measured error rate using the character error rate (CER). CER is the edit distance between the typed text and the reference text divided by the number of characters in the reference. Workers made very few mistakes while copying the text and had an average CER of 0.53%. The text-copy task was completed by 180 unique workers. 53% of the tasks were completed by Indian workers and 44% were completed by US workers. The entry rate for Indian workers was 41 wpm with a 0.72% CER while US workers had an entry rate of 61 wpm with a 0.30% CER. We have included the data collected from both experiments in our sentence metadata.

SUGGESTED TEST SETS

Using the metadata, we created 5 sets of 40 sentences (denoted `mem1`–`mem5`). Each sentence was remembered correctly by 8–10 workers and copied correctly by 8–10 workers. Each sentence had three or more words and was proofread to ensure good grammar. We recommended these sets for evaluations in which memorable text is desired.

In some interfaces user effort may be related to the costs of transitioning between keys (e.g. by Fitts' law). In such cases we might prefer a test set in which combinations of characters appear about as often as in the target domain. Paek and Hsu [5] describe a procedure for creating phrase sets by randomly sampling sets of *n*-grams and choosing the set whose character bigram distribution is closest to the distribution over the entire dataset. We modified their procedure to only select entire sentences since mid-sentence *n*-grams would likely be confusing to participants (e.g. "because it was your"). We created test sets of the 40, 80, 160 and 320 sentences which had a bigram distribution close to the distribution over the entire MOBILE set. We limited sets to sentences with 3–9 words. These test sets (denoted `bi40`–`bi320`) are recommended when a character

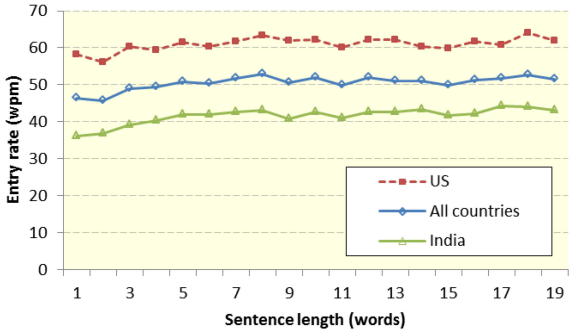

Figure 4: Average entry rate for sentences of different lengths.

Query	Example sentences
8+ words, business messages, 9+ workers had 100% correct memorization	Do you still need me to sign something? I have a high level in my office. Don't make me pull tapes on whether you understood our fee.
4-8 words, personal messages, questions	Is she done yet? Where are they hitting the books at? Could you try ringing her?
Text copied at ≥70 wpm with 0.0% CER	I plan to be in the office tomorrow. I don't think they formally backed out, but effectively that is what has happened. I am at the lake.

Table 4: Example queries using our script and metadata.

distribution representative of mobile email is desired.

From the 200 sentences in our 5 memorable sets, we selected the 40 sentences with a bigram character distribution close to the MOBILE set. This test set (denoted `mem_bi`) is recommended when both memorable text and a representative character distribution is desired.

DISCUSSION

Our dataset is drawn from 44 Enron managers, all using a particular brand of mobile device, and all occurring sometime in the past (circa 2000). While some of the sentences in the dataset were specific to Enron, most were typical of generic business or personal communications. For privacy reasons, the public release of private user emails is rare. We are not aware of any more recent or more diverse sources of actual mobile messages.

Our data was written on BlackBerry devices with a QWERTY thumb keyboard. The sentences do not exhibit the SMS-style texting language or abbreviations that might occur when using a telephone keypad interface. Our test set is not intended to evaluate this style of text entry. It is also unclear how prevalent such abbreviations are today with the rise of touch-screen phones (e.g. iPhone and Android). Even circa 2004, a diary study of 24 users found that 58% of SMS messages were written in unabbreviated form [2].

Our empirical estimation of memorization, entry rates, and error rates were obtained in experiments conducted via crowdsourcing. This resulted in less experimental control than would be possible in a laboratory setting. For example, workers may have completed a varying number of HITs, they may have written sentences down, or they may have taken breaks during timed text entry. However, we were able to collect empirical data for over two thousand sentences, replicated across ten workers. This would have been impractical using a conventional experiment.

Our dataset offers a number of advantages. It is more externally valid than other text sources often used in mobile text entry evaluations. Our sentences were actually written by people in their normal day-to-day mobile email activity. Our dataset has over two thousand sentences providing sufficient data for longitudinal evaluations and for creating subsets that address specific text requirements. Our dataset contains the capitalization, symbols and numbers used by actual mobile users. Thus it enables evaluations using the diverse character sets typical of real-world mobile device usage. However, we have defined subsets that can be used for interfaces with more limited character input capabilities. To aid evaluation of speech recognition interfaces, we have provided normalized versions with verbalized punctuation, numbers and spelling.

The phrase set by MacKenzie and Soukoreff [4] was designed to contain easy to remember text. However, the phrases were not collected from actual mobile messages and the memorability of the phrase set was never verified. In contrast, our dataset is based on genuine mobile emails and provides empirical data regarding sentence memorability. We also provide empirical entry and error rates for each sentence obtained using a full-sized keyboard. Further, we have categorized each sentence as personal, business or Enron specific. Using this metadata, test sets meeting specific requirements can be easily created (see Table 4 for some examples).

CONCLUSIONS

We have described a new resource for mobile text entry researchers: a collection of genuine email sentences written by actual users on mobile devices. We obtained this data by mining sentences written by Enron employees on their BlackBerry mobile devices. The data has been thoroughly inspected, cleaned and organized to make it suitable for use in a variety of text entry evaluations. We have filtered the collection to provide subsets suited for different device capabilities and input modalities. Using a large group of crowdsourced workers, we collected empirical data on how easy sentences in our dataset were to memorize. We also investigated how fast and accurately each sentence could be typed using full-sized keyboards. We have added this additional information as metadata to our dataset. We hope other researchers will find the dataset and results useful when conducting text entry evaluations.

ACKNOWLEDGEMENTS

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

REFERENCES

1. http://www.keithv.com/software/enronmobile/
2. Faulkner, X. and Culwin, F. When fingers do the talking: a study of text messaging. *Interacting with Computers 17*, 2 (2005), 167–185.
3. Klimt, B. and Yang, Y. Introducing the Enron Corpus. *Proc. Email and Anti-Spam* (2004).
4. MacKenzie, I.S. and Soukoreff, R.W. Phrase sets for evaluating text entry techniques. *Ext. Abstracts CHI 2003*, ACM Press (2003), 754–755.
5. Paek, T. and Hsu, B. Sampling representative phrase sets for text entry experiments: a procedure and public resource. *Proc. CHI 2011*, ACM Press (2011), 2477–2480.