Enhancing the Composition Task in Text Entry Studies: Eliciting Difficult Text and Improving Error Rate Calculation

Dylan Gaines
Michigan Technological University
Houghton, MI, USA
dcgaines@mtu.edu

Per Ola Kristensson
University of Cambridge
Cambridge, United Kingdom
pok21@cam.ac.uk

Keith Vertanen
Michigan Technological University
Houghton, MI, USA
vertanen@mtu.edu

ABSTRACT

Participants in text entry studies usually copy phrases or compose novel messages. A composition task mimics actual user behavior and can allow researchers to better understand how a system might perform in reality. A problem with composition is that participants may gravitate towards writing simple text, that is, text containing only common words. Such simple text is insufficient to explore all factors governing a text entry method, such as its error correction features. We contribute to enhancing composition tasks in two ways. First, we show participants can modulate the difficulty of their compositions based on simple instructions. While it took more time to compose difficult messages, they were longer, had more difficult words, and resulted in more use of error correction features. Second, we compare two methods for obtaining a participant's intended text, comparing both methods with a previously proposed crowdsourced judging procedure. We found participant-supplied references were more accurate.

CCS CONCEPTS

• **Human-centered computing** → **Text input**.

KEYWORDS

Text entry; evaluation; composition;

ACM Reference Format:
Dylan Gaines, Per Ola Kristensson, and Keith Vertanen. 2021. Enhancing the Composition Task in Text Entry Studies: Eliciting Difficult Text and Improving Error Rate Calculation. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3411764.3445199

1 INTRODUCTION

When evaluating text entry interfaces, participants are typically asked to transcribe memorable phrases, for example, from the MacKenzie phrase set [4] or the Enron mobile phrase set [7]. An alternative evaluation methodology is to have participants compose novel messages. In this paper, we explore enhancements to composition-based evaluation with a focus on eliciting text that is more likely to require participants make use of an interface's error correction features.

The need for composition-based tasks stems from the fact that transcription tasks are somewhat artificial, since in real-world tasks users will frequently be composing their own thoughts rather than copying existing text [8]. The MacKenzie and Enron phrase sets use short, memorable phrases that average less than six words per phrase [6]. However, in real-world posts made on mobile devices, sentences averaged 11 words [9]. Memorizing sentences this long may be difficult; Vertanen and Kristensson [7] found that sentences of this length were remembered correctly by less than 60% of the sampled participants (n = 386). While the text to be copied can be shown as a reference somewhere in the interface under evaluation, this can also result in unwanted and unrepresentative user behavior, such as the user constantly shifting their attention between the stimulus phrase and the text entry method. It can also be problematic for testing certain interfaces or use scenarios, such as eyes-free text entry methods.

Further creating a need for composition tasks, mobile text language is constantly evolving as new terminology and new texting idioms come in and out of fashion. Composition tasks capture this aspect of the language automatically. In contrast, for transcription tasks such language evolution is more challenging to capture. In practice, phrase sets used for text entry evaluation are, so far, static. Presenting another challenge, Fraco-Salvador and Leiva reported that transcription tasks were found to produce different results when not presented in participants' native language [2]. Composition tasks, on the other hand, are able to be performed in any language supported by the input method, without requiring the translation of phrase sets.

A related challenge is carrying out robust analysis for compositions that lack a definitive reference text. The lack of reference text make it difficult to calculate metrics, such as error rate. All these factors motivate further work on fine-tuning effective composition tasks for text entry evaluation.

We make two contributions to enhance composition tasks. First, we investigate the feasibility of a simple method for eliciting more difficult text from participants. We anticipate this may be used in order to better test the effectiveness of an interface's error correction or error avoidance features. We find participants can successfully modulate the difficulty of their text and that difficult compositions affect writing time and the use of error correction features. Second, we compare several methods for obtaining a participant's intended text. These methods can be utilized in the calculation of error rates used to create comparisons between different interfaces in composition-based experiments. We find participant-supplied references are more accurate compared to a crowdsourced judging procedure.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445199

Figure 1: Instructions in Study 1 for Easy (left), Hard (middle), and post-input feedback screen (right).

2 APPROACH

To elicit more challenging compositions, we developed instructions asking participants to compose things they thought would or would not cause recognition errors (Figure 1). As we will see, our instructions successfully changed participants' compositions. In Study 1, we had participants compose easy and hard compositions in separate conditions. In Study 2, we interleaved easy and hard tasks at random in a single study condition.

One method for obtaining a participant's intended text for use in calculating error rate is crowdsourcing. In a few instances, Vertanen et al. [5, 8] had Amazon Mechanical Turk workers judge compositions to determine the reference text. Although this method allows for an approximation of error rate, it may not always be accurate, especially in cases where some of a participant's intended words are less common. Previous work also suggests that even for relatively easy text, crowdsourced judging may underestimate the true error rate [5]. A similar method was used by Karat et al. [3], who asked peers to count errors in final compositions and evaluate the overall message clarity. Another method, used by Arnold et al. [1], calculated the number of backspaces a user performed and divided by the total number of taps. While this metric can be helpful in determining the initial accuracy of users, it does not capture uncorrected errors.

We propose two alternative methods, extending the composition procedure described by Vertanen and Kristensson [8]. In Study 1, participants first invented a composition and typed it on a smartwatch keyboard. After each task, they also typed their intended text on a laptop. Participants could see what they typed on the laptop and make corrections by backspacing. Of course, typing even on a desktop keyboard is subject to mistakes. However, these mistakes are likely to be more minor than those of a recognition-based input method. The text typed on the desktop keyboard can be corrected by a crowdsourced protocol. Alternatively, a simpler approach, and the one we take here, is to have an experimenter review the text and correct obvious typographical errors. This corrected text serves as the reference text.

In Study 2, after typing on the watch, participants dictated their text to the experimenter. The experimenter typed it on a desktop computer, clarifying any hard words. To minimize recall problems, participants specified their intended text immediately after each composition. We also independently obtained a reference for each composition in both studies via a crowdsourced procedure [8]. This was done to further validate whether crowdsourced judging underestimates the error rate as has been previously shown when using a known reference in a text copy task [5]. Of course in an actual study of a text input method, only one method of obtaining a reference would be needed. Our goal here was to explore the trade-offs in how the reference is obtained:

Figure 2: Study 1 smartwatch interface. Users type each letter in a word before swiping right to obtain the most likely recognition. The interface shows the nearest key label to a user's touch. After 500 ms, the label changes color to signify the letter is locked and no longer subject to auto-correct.

- **Crowdsourcing** Using crowdsourcing demands no additional time or effort from the participants. However, as previously discussed, it may underestimate the true error rate. It also means the experimenter must manage crowd work on a microtask market such as Amazon Mechnanical Turk.
- **Laptop** The participant types the transcript on a laptop. This demands additional time and effort from the participant and may require switching devices. The experimenter may need to later correct any typographical errors in the participant's laptop text.
- **Dictate** The participant speaks the reference and the experimenter enters it. This demands some participant time and effort. The experimenter can clarify the intended text immediately. This method requires the experimenter closely interact with the participant throughout the study.

3 STUDY 1: EASY AND HARD INSTRUCTIONS

3.1 Interface

Participants entered text on a smartwatch keyboard containing the characters A-Z and apostrophe. The keyboard measured 29 mm × 13 mm on the 29 mm × 29 mm screen of a Sony Smartwatch 3. Participants could indicate a space by swiping to the right, or a backspace by swiping to the left. Indicating a space would also trigger recognition by a decoder based on the VelociTap decoder [10]. When a participant's finger was in contact with the keyboard the closest letter would be shown over the top of the text area (Figure 2 left). If a participant held their finger on the screen for 500 ms, that letter would be highlighted orange and locked (Figure 2 right), which prevented it from being changed by the decoder. Previously recognized words in the composition as well as the letters closest to each tap in the current word were displayed in the area above the keyboard.

3.2 Method

The goal of the first study was to see whether we could elicit more challenging text in a composition-style task. This was a within-subject experiment with two counterbalanced conditions. In the

Easy condition, participants were asked to invent a message they thought would be be recognized with no errors (Figure 1 left). In the Hard condition, participants were asked to invent a message they thought would be be recognized with one or more errors (Figure 1 middle).

After each composition, participants were asked to type their intended text on a laptop (Figure 1 right). We manually reviewed the laptop and corresponding watch recognition results correcting obvious laptop typing mistakes. We corrected 8 out of the 320 total compositions. The corrected laptop entries were taken as the reference transcripts for error rate calculation.

16 participants completed this study immediately after Experiment 1 in the paper by Vertanen et al. [6]. This prior experiment allowed participants to gain familiarity with the watch text entry interface. Our studies here serve both to investigate composition methodology and to see the correction behavior exhibited in practice. Participants were 18–27 years old (mean 19.1) and 10 identified as male, 2 identified as female, and the rest chose to not answer. All users were enrolled at a university and rated the statement "I consider myself a fluent speaker of English" a 7 on a 7-point Likert scale where 7 was strongly agree. They were paid $10 to take part in a one hour session. 10 users reported never using a smartwatch before, while 3 reported using one frequently and 3 occasionally.

In this prior experiment, users transcribed a mix of phrases that either were completely in-vocabulary or had an out-of-vocabulary (OOV) word. Users completed one condition with the letter-locking feature on and one condition where it was off. This prior experiment familiarized users with the interface and the error correction feature. Our followup study reported here aimed to 1) measure their use of letter locking in a more naturalistic composition task, and 2) investigate if users would be willing and able to invent compositions that stimulated use of letter locking similar to how transcribing OOV phrases did. In the study here, participants could lock letters in both conditions. Participants did two practice compositions followed by ten compositions in each condition. We did not analyze the practice tasks.

We measured *entry rate* in words-per-minute (wpm), with a word being five characters including space. We calculated entry time from a participant's first tap until the last recognition or correction was made. We measured *error rate* using Character Error Rate (CER). CER is calculated by dividing the edit distance between a participant's final text and the reference text by the number of reference characters. We measured *task time* by dividing the time spent in a condition by the number of tasks. This includes the time for thinking of a composition, typing on the watch, and providing the intended text. We measured *backspaces per character* by dividing the number of backspaces performed by the user by the final number of output characters.

We report a number of metrics about what participants wrote. We measured the *characters per composition* (including spaces) and the average *characters per word*. We report the per character *perplexity*. Perplexity measures the average number of choices the recognizer has when predicting the next character using its language model. For example, a language consisting of the digits 0–9 with each digit being equally probable has a perplexity of 10. Text with less common words typically has a higher perplexity. We calculated the *OOV rate* as the percentage of words that were out-of-vocabulary (OOV) with respect to the 100 K vocabulary used by our recognizer.

We compared our participant-supplied references with the crowdsourced protocol from Vertanen and Kristensson [8]. We asked workers on Amazon Mechanical Turk to correct each composition. As in [8], if a worker thought a sentence was completely correct, its CER was taken as 0%. If a worker thought a sentence was not correctable, it was taken as 100%. Otherwise its CER was calculated based on a worker's provided correction. The *judged CER* was the median of the workers' error rates. Each worker received 30 compositions, 10 of which had known corrections. Workers received a random mix of easy and hard compositions. We only kept workers who got 60% of the known corrections exactly correct (including case and punctuation).

3.3 Results

Figure 3 shows the main results. Table 1 provides numeric results and statistical tests. A Shapiro-Wilk tests found the difference of paired samples deviated from normal for letter lock percentage ($W = 0.75$, $p < .001$), characters per composition ($W = 0.88$, $p < .05$), characters per word ($W = 0.77$, $p < 0.005$), and perplexity ($W = 0.49$, $p < .001$). For metrics that violated normality, we used a Wilcoxon signed-rank test. All other tests used a dependent t-test.

Participants were slower entering more difficult compositions: 19.7 wpm in Hard versus 24.7 wpm in Easy. The slower speed may be due to their increased use of letter locking: 8.8% in Hard versus 0.8% in Easy. Another explanation is that entry was slowed by users stopping mid-composition to think of difficult things to write. This may be indicated by the substantially longer task time of 49.4 s in Hard versus 32.3 s in Easy. All differences were statistically significant (Table 1).

For comparison, we found that the task time in the Lock condition of Experiment 1 in the work done by Vertanen et al. [6] was 23.9 s ± 2.3 (95% CI) with participants spending 74% of their time typing on the watch. The remaining 26% constitutes overheads in memorizing the phrases they were copying. Even though participants in our composition study had more practice with the interface, Easy compositions took around 8 s longer per task. In Easy, participants spent 42% of their time typing on the smartwatch and 22% typing their intended text on the laptop. The remaining 36% constitutes overheads associated with thinking of what to write or switching between the watch and laptop. While the composition task did result in some experimental overhead, the additional time required of participants was modest.

Using the reference text typed by the participant (and possibly corrected by the experimenter), we found the error rate was elevated at 5.4% in Hard versus 3.7% in Easy. However, this difference was not statistically significant (Table 1). The composition error rate in Easy was similar to the 3.3% error rate reported in Experiment 1 of [6] in which participants copied memorable phrases in the Lock condition. To measure errors corrected by participants, we also calculated backspaces per final output character. While slightly higher in Hard, this difference was not significant.

After dropping inaccurate workers, 318 of the compositions were judged by five or more workers while the remaining two compositions had only three workers. Compared to past work [5], we had

Metric	Easy		Hard		Statistical test details		
Entry rate (wpm)	24.7 ± 3.6	[16.5, 36.1]	19.7 ± 3.6	[11.8, 32.5]	$t(15) = 5.95$	$r = 0.84$	$p < .001$
Error rate (CER %)	3.7 ± 2.0	[0.0, 11.7]	5.4 ± 2.6	[0.7, 17.2]	$t(15) = -1.46$	$r = 0.35$	$p = .16$
Task time (s)	32.3 ± 4.5	[22.4, 50.6]	49.4 ± 9.7	[25.8, 90.4]	$t(15) = -6.29$	$r = 0.85$	$p < .001$
Letter lock (%)	0.8 ± 0.7	[0.0, 4.5]	8.8 ± 6.6	[0.0, 40.3]	Wilcoxon	$r = 0.72$	$p < .005$
Backspaces per char	0.065 ± 0.04	[0.0, 0.2]	0.080 ± 0.05	[0.0, 0.3]	$t(15) = -1.25$	$r = 0.31$	$p = .23$
Chars per composition	25.2 ± 5.4	[8.8, 44.4]	32.2 ± 7.8	[10.9, 65.6]	Wilcoxon	$r = 0.69$	$p < .005$
Chars per word	4.6 ± 0.2	[4.1, 5.7]	5.3 ± 0.3	[4.1, 6.8]	Wilcoxon	$r = 0.67$	$p < .005$
Perplexity	6.6 ± 2.9	[3.3, 25.7]	13.0 ± 9.9	[5.0, 79.4]	Wilcoxon	$r = 0.81$	$p < .001$
OOV rate (%)	2.4 ± 4.2	[0.0, 30.0]	13.3 ± 5.9	[3.2, 45.8]	$t(15) = -6.20$	$r = 0.85$	$p < .001$

Table 1: Study 1 results. Participants composed easy or hard text. Results format: mean ± 95% CI [min, max].

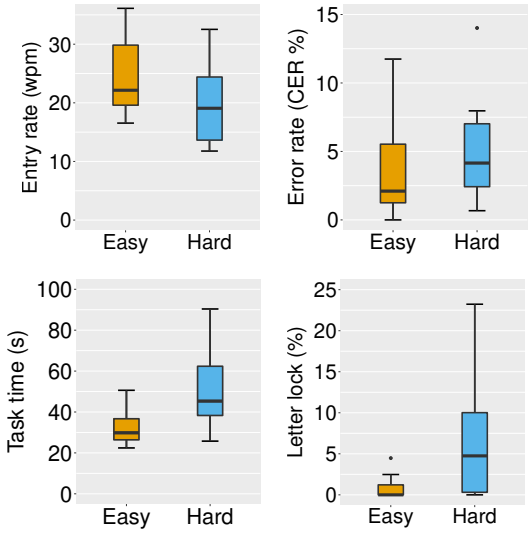

Figure 3: Entry rate, error rate, task time, and letter lock percentage in Study 1.

Condition	Composition
Easy	there's a stray that used to hang around
	i broke my screen lol
	cinnamon is pure evil
	the leaves are already falling here
	finally done with the easy one
Hard	bbc is the best broadcast station
	bacchus is the god of wine
	the uss zumwalt is a new class of destroyer
	the taj mahal is amazing
	lukerdoo is very protective of the nachos

Table 2: Example compositions from Study 1.

to lower the bar for considering workers as accurate. We also had to repeatedly launch the Amazon task on subsets of the compositions to arrive at sufficient judgements across all compositions. We found a number of workers were simply judging all 30 compositions as correct or uncorrectable. While crowdsourced judging may be useful in cases where a reference is difficult or impossible to record during an experiment, it may also require extra effort to conduct.

We found a judged CER of 3.8% in Hard and 2.7% in Easy. The relative difference in judged CER matches what we found with the reference text provided by the participants. However, similar to [5], we found the judged CER tended to underestimate the true error rate of the compositions.

As we expected, the hard compositions seemed more difficult for crowdsourced workers to judge. Workers spent on average 30 s on the Hard compositions versus 22 s on the Easy compositions. Workers judged Hard compositions as impossible to correct 22% of the time and completely correct 28% of the time versus 14% and 32% for the Easy compositions.

In both conditions, participants invented plausible compositions (Table 2). As evidenced by the increased out of vocabulary rate and use of error correction features in Hard, participants seemed to invent compositions that were indeed harder to recognize. However, inventing hard compositions took substantial additional time, on average tasks in Hard took 48% more time then Easy. Further, Hard compositions took twice as long as the transcription tasks reported in Vertanen et al. [6].

Compositions tended to be shorter at 25 characters in Easy compared to 32 characters in Hard (Table 1 bottom). Participants wrote slightly shorter words of 4.6 characters in Easy compared to 5.3 characters in Hard. The perplexity under our recognizer's character language model was lower at 6.6 in Easy versus 13.0 in Hard. In participants' Easy compositions 2.4% of words were OOV compared to 13.3% in Hard. All differences were statistically significant. Based on these metrics, our instructions were successful at eliciting more difficult compositions including OOV words.

Additionally, we measured a few error correction metrics. Though as we mentioned earlier there was no significant difference in the backspaces per character, there was a significant difference in the participants' use of the letter lock feature. Participants used a long tap to lock a letter, preventing the recognizer from changing a character, 0.8% of the time in the Easy condition, and 8.8% of the time in the Hard condition. This suggests that participants were less confident that the recognizer would be able to accurately determine their intended text in the Hard condition.

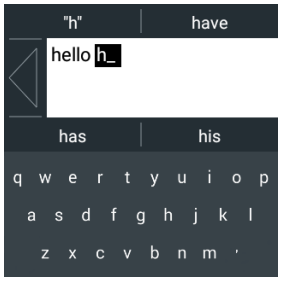

Figure 4: Study 2 smartwatch interface. The center highlighted text is the best word assuming typing is complete for this word. The top left slot shows the literal characters typed. The other slots show word completions. The left arrow is a backspace key. Options are selected either by tapping or by a swipe gesture (e.g. up-and-left for the literal slot).

4 STUDY 2: MIXING EASY AND HARD TASKS

In Study 1, we had explicit conditions in which participants were asked to compose easy or hard messages. In some cases, it might be desirable to mix tasks together in order to investigate how an interface performs for varying difficulties of text all within the same experimental condition. Mixing distinct easy and hard tasks in this way makes it simple to analyze performance on the input of easy and more challenging text separately. Additionally, in Study 1 we sometimes found it difficult after the study to correct typos participants made on the laptop. Having participants type the reference on a laptop may be also difficult in some studies, e.g. text input while wearing a virtual reality head-mounted display. Study 2 used the same easy and hard composition instructions as in Study 1, but mixed them together at random. The instructions were to invent a message having either no auto-correct errors (termed easy) or one or more auto-correct errors (termed hard).

4.1 Interface

The interface used in Study 2 was the same as in Study 1 with the following modifications. The keyboard had five word predictions and a backspace key as shown in Figure 4. One of the five predictions was always the literal keys typed by the participant. If the participant selected this *literal* prediction slot, the text was not subject to auto-correction. These interface elements could be triggered by either tapping them or swiping in the direction of the slot. To backspace, participants could either press the key or swipe left as in Study 1. As in Study 1, participants could long press to lock individual letters to avoid potential auto-correct errors.

4.2 Method

Participants completed four practice compositions followed by ten evaluation compositions. The practice and evaluation had an equal number of easy and hard tasks. Due to a bug, one participant received six hard and four easy tasks, and four participants received four hard and six easy tasks. We restored balancing for these participants by using only the first four tasks of each type (eight total). We used all ten tasks for the remainder of the participants.

After each composition, participants were shown a screen instructing them to dictate their intended text to the experimenter. The experimenter typed the dictated text into a desktop computer at the desk where the participant was seated. The experimenter verified any difficult words with the participant. This text will serve as the reference text for error rate calculation.

24 participants completed this study immediately after completing Experiment 3 in the paper by Vertanen et al. [6]. Similar to Study 1, completing this prior experiment allowed participants to gain familiarity with the interface. The focus of the study here is to investigate composition-based evaluation and see the correction behavior exhibited in practice. None of these participants were involved in Study 1. They were 18–22 years old and 15 identified as male, while 6 identified as female and the rest chose to not answer. All users were enrolled at a university. When asked to rate the statement "I consider myself a fluent speaker of English" on a 7-point Likert scale where 7 is strongly agree, 2 users answered 6 and the remainder answered 7. 13 users reported never using a smartwatch before, while 3 reported using one all the time. The remainder reported using one occasionally. The participants took part in a one-hour session and were paid $10.

In the prior experiment [6], participants typed memorable phrases into the smartwatch keyboard. This prior experiment was designed to investigate differences between the method of selecting prediction slots (i.e. tap or swipe gestures) and showed both methods had similar entry and error rates. There was however an increase in error rate for OOV phrases over in-vocabulary phrases across all conditions. Vertanen et al. [6] conjectured that this may have been due to users having difficulty remembering harder phrases. By using composition tasks instead of transcription tasks in this study, we can eliminate this memory factor.

The study reported here involved just a single condition in which participants were free to either swipe or tap to select word predictions or to trigger a backspace. From the standpoint of the prior research, this condition allowed observing how participants interacted when given a free-choice in a more naturalistic input setting. In this paper, we focused on whether interleaving easy and hard composition tasks resulted in measurably different compositions, higher usage of correction features, and on how well the verbal dictation procedure worked.

4.3 Results

Table 3 shows the results for Study 2 along with statistical tests. Shapiro-Wilk tests found the difference of paired samples deviated from normal for the error rate ($W = 0.83$, $p < .005$), letter lock percentage ($W = 0.58$, $p < .001$), characters per composition ($W = 0.85$, $p < .005$), and perplexity ($W = 0.63$, $p < .001$). For metrics that violated normality, we used a Wilcoxon signed-rank test. All other tests used a dependent t-test.

As shown in Table 3, participants were once again slower typing hard compositions (16.7 wpm) as opposed to easy compositions (20.4 wpm). This difference was statistically significant. This is consistent with what we found in Study 1. As in Study 1, this may have been due to increased letter locking (3.7% in Hard vs. 0.4% in Easy) or by time spent thinking of difficult words to type. For comparison, Experiment 3 in Vertanen et al. [6] showed users typed at

Metric	Easy		Hard		Statistical test details		
Entry rate (wpm)	20.4 ± 2.4	[10.4, 33.8]	16.7 ± 2.9	[7.6, 29.9]	$t(23) = 3.14$	$r = 0.55$	$p < 0.01$
Error rate (CER %)	0.7 ± 0.5	[0.0, 4.4]	1.5 ± 1.0	[0.0, 8.3]	Wilcoxon	$r = 0.44$	$p < 0.05$
Task time (s)	47.0 ± 7.7	[24.6, 108.6]	65.4 ± 10.8	[31.3, 128.8]	$t(23) = -3.27$	$r = 0.56$	$p < 0.01$
Letter lock (%)	0.4 ± 0.5	[0.0, 5.1]	3.7 ± 3.0	[0.0, 25.6]	Wilcoxon	$r = 0.72$	$p < .001$
Backspaces per char	0.13 ± 0.07	[0.0, 0.8]	0.16 ± 0.07	[0.0, 0.6]	$t(23) = -1.15$	$r = 0.23$	$p = 0.26$
Literal slot (%)	4.7 ± 3.1	[0.0, 25.0]	8.6 ± 4.8	[0.0, 46.7]	$t(23) = -2.07$	$r = 0.40$	$p < .05$
Chars per composition	29.7 ± 6.4	[13.2, 65.2]	34.3 ± 5.4	[18.2, 66.0]	Wilcoxon	$r = 0.46$	$p < .05$
Chars per word	4.8 ± 0.2	[4.2, 5.7]	5.2 ± 0.3	[4.4, 6.4]	$t(23) = -3.19$	$r = 0.55$	$p < .005$
Perplexity	5.5 ± 1.5	[3.0, 15.5]	8.8 ± 5.7	[3.3, 54.1]	Wilcoxon	$r = 0.40$	$p = .053$
OOV rate (%)	1.7 ± 1.9	[0.0, 11.8]	9.4 ± 5.1	[0.0, 29.2]	$t(23) = -3.75$	$r = 0.62$	$p < .005$

Table 3: Study 2 results. Participants composed easy or hard text. Results format: mean ± 95% CI [min, max].

13.9 wpm when transcribing phrases with out-of-vocabulary words and 21.3 wpm for completely in-vocabulary phrases.

Consistent with Study 1, the task time for hard compositions was higher than for easy compositions, 65.4 s compared to 47.0 s. In this study, participants spent 47.3% and 45.6% (in Hard and Easy, respectively) of the task time typing on the smartwatch, while the remainder consisted of the overhead of both thinking of compositions and dictating their intended text to the experimenter.

When writing hard compositions, participants selected the literal prediction slot for 8.6% of words compared to 4.7% for easy compositions. This difference was significant. This once again shows that the hard instructions were successful at causing more use of the interface's error correction features.

Participants also had a significant difference in character error rates between easy and hard composition tasks. On average, participants' easy compositions had a 0.7% CER, while their hard compositions had a 1.5% CER (Table 3). As in Study 1, we asked workers on Amazon Mechanical Turk to correct the compositions. Again we had to run several iterations of the judging to arrive at sufficient judgements by workers who answered 60% of the known corrections correctly. The judged CER was 0.6% for easy compositions and 0.8% for hard compositions. As in Study 1, judged CER seemed to underestimate the error rate compared to the participant-provided references. For corrected errors, the backspaces per final output character was slightly higher in Hard, but similar to Study 1, this was not statistically significant (Table 3).

In Experiment 3 of the paper by Vertanen et al. [6], 38% participants reported that they preferred selecting suggestion slots with swipe gestures, while 29% preferred tap gestures. The remainder preferred the Hybrid condition. Given all options in our composition-based study, participants used a tap gesture 74.6% of the time and used a swipe gesture only 25.4% of the time. Furthermore, only 6 of the 24 participants used swipe gestures more than tap gestures.

We found similar trends to Study 1 when analyzing the text of participants' compositions (Table 3 bottom). Compositions were shorter in Easy at 30 characters versus Hard at 34 characters. Words were shorter in Easy at 4.8 characters per word versus 5.2 in Hard. In participants' Easy compositions 1.7% of words were out-of-vocabulary compared to 9.4% in Hard. All these differences were statistically significant. While perplexity was again lower for Easy compositions at 5.5 versus 8.8 in Hard, this difference was not significant.

While the generation of a phrase set was not the focus of this paper, we have released the compositions from Studies 1 and 2 for use in future studies that require challenging phrases for conducting traditional transcription-based evaluations. They are included as part of the supplementary material in the ACM Digital Library.

5 DISCUSSION

The first goal of this work was to investigate whether participants could compose messages on-demand that can challenge a text entry method with strong auto-correction capabilities. It was not obvious from the onset whether participants would really do this as it requires participants to voluntarily make their input process more time-consuming and challenging. However, we found that participants are indeed capable and willing to modulate the difficulty of their text. This was evidenced by an increase in composition length, characters per word, and OOV rate in their compositions in both studies. We also found participants made more use of the lock letter error avoidance feature in both studies. At least for users with significant experience with an auto-correcting keyboard, eliciting challenging text seems to be as easy as just asking participants to invent things they anticipate will be problematic. This result may not hold true for users that are not as experienced with using an auto-correcting keyboard.

It is important to note the compositions observed in this study may differ from real-world compositions for a couple of reasons. First, our composition tasks were conducted immediately after transcription tasks. In general, this will not be the case for all studies making use of composition tasks. The goal of the studies here was to determine if users could modulate the difficulty of their text to exercise error correction features, not to elicit as realistic of compositions as possible. Another possible concern is that prompted composition does not accurately represent real-world composition. However, real-world composition usually occurs in response to some sort of stimulus. This stimulus could be a text message or a thought that a person wants to make a note of. The prompts given in these studies are simply another, albeit slightly more open-ended, form of this stimulus. We think this reflects how users may initiate

	MacKenzie phrases	Enron phrases	Exp 2 [8]	Exp 2 [5]	Study 1 Easy	Study 1 Hard	Study 2 Easy	Study 2 Hard
Words per phrase	5.43	5.31	6.92	5.15	5.51	5.99	6.09	6.56
Chars per phrase	28.63	25.06	32.22	23.78	25.21	32.18	29.40	34.26
Chars per word	5.38	4.70	4.63	4.63	4.61	5.29	4.80	5.19
Perplexity	4.60	4.25	4.50	4.53	6.60	12.99	5.59	8.47
OOV rate (%)	0.08	0.09	0.31	0.60	2.42	13.32	1.81	9.24

Table 4: Text complexity in two transcription phrase sets, two prior composition studies, and our composition studies. Results are the mean of all phrases and not the mean of participant means (as in Tables 1 and 3).

a text messaging conversation, for example, but may differ from writing replies in an existing conversation.

So how did our participants' compositions compare to past work? To measure the complexity of compositions written without explicit instructions about creating challenging text, we analyzed the compositions of 46 US Amazon workers in Experiment 2 of [8] (448 compositions) and 24 users in Experiment 2 of [5] (249 compositions). We also analyzed two standard transcription phrases sets: the MacKenzie phrase set [4] (500 phrases), and the Enron memorable set [7] (189 phrases).

As shown in Table 4, participants in these previous composition studies composed messages of roughly similar length to the MacKenzie and Enron phrases. Notably participants in [8] created longer messages, probably as a result of likely using a desktop keyboard. Participants in [5] seemed to generate slightly shorter messages, perhaps as a result of using a watch keyboard. Both previous composition studies and phrase sets had similar perplexities. However the OOV rate was markedly higher for the composition studies, showing that even without explicitly asking for challenging text, composition may encourage writing with a richer vocabulary. This demonstrates that participants do tend to gravitate towards simpler text when given the instructions from Vertanen et al. [8].

Table 4 shows that our Easy conditions generated text that was similar in length to these prior studies and phrase sets. Perplexity and OOV rates were slightly elevated in our Easy conditions compared to prior work. We conjecture this could be due to the error correction features present in our watch interface encouraging more ambitious writing. In our Hard conditions, we see longer compositions with much higher perplexities and OOV rates. Interestingly, our Study 2 that mixed easy and hard composition tasks together may have somewhat lowered perplexity and OOV rate compared to Study 1 where participants did all hard compositions in a single block. It could be that mixing tasks together hinders participants from getting into a challenging text writing "flow". This would need further study to validate, but our results do show mixing easy and hard tasks was able to elevate composition complexity markedly compared to previous composition studies.

Our second goal was to expand the ways text entry researchers can administer a composition-based text entry study by teasing out the positive and negative factors of having participants provide reference transcripts compared to using crowdsourced judging [8]. Our two proposed procedures for performing composition tasks did allow for a more accurate calculation of error rate compared to previous work. In Study 1, we instructed participants to enter their intended text on a laptop computer and found that the actual error rates were higher than those calculated using crowdsourced judging. However, this procedure may not be feasible in all situations, such as when the experimental interface is in virtual reality and the reference is being typed on a physical keyboard. This would require the removal of the head-mounted display after each task. Participants may also make typos that may be difficult to correct later.

In Study 2, participants dictated their intended text to the experimenter, and found a similar difference compared to crowdsourced judging. Although the average task time was longer than the procedure in Study 1, with only a marginally higher percentage of time spent typing (46.7% compared to 42% in Study 1), this procedure allows for back-and-forth verification of intended spelling between the participant and the experimenter. For this reason it is arguably more accurate, though it does require constant oversight by the experimenter. Both the laptop and dictation procedures avoid the experimenter needing to deal with the extra hassle and expense of crowdsourced judging.

This paper furthers our understanding of the composition task in text entry, which has higher external validity than a transcription task, at the cost of lower internal validity. A previous set of studies on the composition task has alleviated several internal validity concerns, such as 1) compositions being too slow and with too much variation; 2) the possible interference of cognitive overhead in composition resulting in unacceptable increases in variance between participants; 3) too much of a participant's time is spent planning compositions rather than writing; 4) the lack of reference text may make calculating error rate problematic; and 5) participants may lack sufficient imagination to generate compositions [8]. This work improves the composition task further by 1) demonstrating that participants can in fact modulate the difficulty of their text; and 2) teasing out the trade-offs between different methods for arriving at an accurate transcript of the composition.

6 CONCLUSIONS

When investigating the error-correcting capabilities of a text entry interface, it is important that participants actually exercise those capabilities. Although this can be artificially introduced by having users transcribe difficult phrases (e.g. as in [6]), phrase sets may not always be up-to-date as language evolves. Further, a composition-style task allows researchers to better understand how a text entry system performs in real-world use. This paper showed participants can compose messages that are more difficult for an auto-correcting

keyboard, as demonstrated by the significantly higher use of error correction features in both studies when participants were instructed to compose hard messages compared to easy messages. In addition, we have investigated the trade-offs inherent in different methods for obtaining accurate transcripts for compositions. We hope this work will help stimulate further use of composition tasks in text entry studies and in particular assist text entry researchers in situations when transcription tasks are impractical or unsuitable.

ACKNOWLEDGMENTS

This work was supported by Google Faculty awards (K.V. and P.O.K.), EPSRC grants EP/N010558/1, EP/N014278/1, EP/R004471/1 (P.O.K), NSF IIS-1909248 (K.V. and D.G.), and an NSF Graduate Research Fellowship 2034833 (D.G).

REFERENCES

[1] Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. 2016. On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. ACM, New York, NY, USA, 603–608. https://doi.org/10.1145/2984511.2984584

[2] Marc Franco-Salvador and Luis A. Leiva. 2018. Multilingual phrase sampling for text entry evaluations. *International Journal of Human-Computer Studies* 113 (2018), 15 – 31. https://doi.org/10.1016/j.ijhcs.2018.01.006

[3] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. ACM, New York, NY, USA, 568–575. https://doi.org/10.1145/302979.303160

[4] I. Scott MacKenzie and R. William Soukoreff. 2003. Phrase Sets for Evaluating Text Entry Techniques. In *Extended Abstracts on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI EA '03)*. ACM, New York, NY, USA, 754–755. https://doi.org/10.1145/765891.765971

[5] Keith Vertanen, Crystal Fletcher, Dylan Gaines, Jacob Gould, and Per Ola Kristensson. 2018. The Impact of Word, Multiple Word, and Sentence Input on Virtual Keyboard Decoding Performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 626, 12 pages. https://doi.org/10.1145/3173574.3174200

[6] Keith Vertanen, Dylan Gaines, Crystal Fletcher, Alex M. Stanage, Robbie Watling, and Per Ola Kristensson. 2019. VelociWatch: Designing and Evaluating a Virtual Keyboard for the Input of Challenging Text. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 591, 14 pages. https://doi.org/10.1145/3290605.3300821

[7] Keith Vertanen and Per Ola Kristensson. 2011. A Versatile Dataset for Text Entry Evaluations Based on Genuine Mobile Emails. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices & Services* (Stockholm, Sweden) *(MobileHCI '11)*. ACM, New York, NY, USA, 295–298. https://doi.org/10.1145/2037373.2037418

[8] Keith Vertanen and Per Ola Kristensson. 2014. Complementing Text Entry Evaluations with a Composition Task. *ACM Transactions of Computer Human Interaction* 21, 2, Article 8 (February 2014), 33 pages. https://doi.org/10.1145/2555691

[9] Keith Vertanen and Per Ola Kristensson. 2019. Mining, Analyzing, and Modeling Text Written on Mobile Devices. *Natural Language Engineering* (2019), 1–33. https://doi.org/10.1017/S1351324919000548

[10] Keith Vertanen, Haythem Memmi, Justin Emge, Shyam Reyal, and Per Ola Kristensson. 2015. VelociTap: Investigating Fast Mobile Text Entry Using Sentence-Based Decoding of Touchscreen Keyboard Input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 659–668. https://doi.org/10.1145/2702123.2702135