Don't Look Up: The Cost of Attention to Stimulus Phrases in Mobile Text Entry Evaluations

Andreas Komninos
University of Patras
Computer Engineering & Informatics Dept.
Patras, Greece
akomninos@ceid.upatras.gr

Angeliki Tsiouma
University of Patras
Computer Engineering & Informatics Dept.
Patras, Greece
tsiouma@ceid.upatras.gr

Georgia Gogoulou
University of Patras
Computer Engineering & Informatics Dept.
Patras, Greece
ggogoulou@ceid.upatras.gr

John Garofalakis
University of Patras
Computer Engineering & Informatics Dept.
Patras, Greece
garofala@ceid.upatras.gr

ABSTRACT

Transcription tasks have been long used as the de-facto evaluation method in mobile text entry research. Evaluations use memorable phrase sets, in order to prevent participants from devoting more attention to the stimulus phrase than the bare minimum. We present evidence from an eye-tracking study, demonstrating that the attention devoted to the stimulus phrase is much higher than might be expected. In fact, attention to the stimulus phrase takes up almost 50% of participant attention spent outside the keyboard area, and overall 25% of participant attention throughout any single transcription task. We explore a modification to the transcription task aimed at reducing this level of visual attention, without finding any statistically significant differences. These findings raise important questions on the continued use of the transcription task as the mainstream evaluation method for mobile text entry research.

CCS CONCEPTS

• **Human-centered computing** → **Keyboards**; **Text input**; *User studies*; **Mobile devices**; **Touch screens**.

KEYWORDS

Mobile text entry, Eye tracking, Transcription tasks, Evaluation

ACM Reference Format:
Andreas Komninos, Angeliki Tsiouma, Georgia Gogoulou, and John Garofalakis. 2022. Don't Look Up: The Cost of Attention to Stimulus Phrases in Mobile Text Entry Evaluations. In *26th Pan-Hellenic Conference on Informatics (PCI 2022), November 25–27, 2022, Athens, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3575879.3576015

1 INTRODUCTION

Transcription tasks in mobile text entry research are the de-facto evaluation method in both laboratory and field-based experimental settings, asking participants to type phrases which are presented to them on the screen of the device used in the experiment. It has been observed that generally, during transcription tasks, participants often check for (and if detected, correct) errors after typing just a few characters, leaving very few errors uncorrected. This behaviour contrasts common experience of real-world text entry, where both uncorrected errors, or even inappropriate auto-corrections frequently go unnoticed. This discrepancy is problematic for lab evaluation of novel text entry methods which aim to offer support for error detection and correction. To collect sufficient data, participants must enter large numbers of phrases, risking fatigue and disengagement from the task. An alternative is to employ field studies, but they are harder to recruit for, and burdened with internal validity issues. It is also not known whether the persistent display of the phrase to be transcribed affects the observed error rates.

In this paper, we quantify the division of participants' attention while performing text entry, during a transcription task. Further, we designed a modification to the text entry task, which aimed to reduce error-checking behaviour by participants. Even though this modification did not prove to be successful, the insights on the attention costs inherent to the transcription task are a novel and important breakthrough, raising concern, and prompting a call to re-think its suitability for lab-based text entry research.

2 RELATED WORK

Lab studies in text entry use transcription tasks as the de-facto method [20]. Participants are asked to "quickly and accurately" copy short phrases that appear on the device used for entry, which are selected from sets that have been validated for memorability (e.g. [13, 18]). Participants are assumed to use these phrases as a prompt at the start of the task, and then enter text without having to look at the stimulus phrases again, or, at least, not often.[14, 20].

Transcription tasks have been criticised for low external validity (unrealistic) [20]. They have also been found to significantly alter natural user error-making behaviour in the lab [5], therefore restricting the ability of researchers to evaluate text entry method improvements focused on supporting error detection and correction. Input errors during studies are so rare, that researchers resorted to various protocol modifications to elicit more. Some artificially injected errors into the user's input stream by covertly replacing typed characters [1, 9, 12, 16]. Another method has been to present

PCI 2022, November 25–27, 2022, Athens, Greece

the user with a pre-typed phrase which includes errors, and to ask users only to fix those errors [12].

User aversion to errors leads users to type more slowly, in order to avoid the cognitive cost of detecting and correcting them [4]. Less is known about how this cost is avoided, in terms of attention paid to each of the three display areas of interest in a transition task (the keyboard, stimulus phrase and text entry field). Evidence from the two existing studies using eye-tracking during mobile transcription, show that participants spend considerable time gazing back and forth between the keyboard, and the two screen areas associated with the task [6, 8]. However, in both studies, the stimulus phrase and typing area were positioned adjacently, therefore it is not possible to ascertain how attention is divided between the two.

Is there some inherent aspect of the transcription task that could be causing high attention to the stimulus phrase area, thus affecting the observability of error-making and error-correcting behaviour during experiments? If such a phenomenon could be quantified, a possible cause might be that the assumed memorability of stimulus phrases is incorrect, despite use of "memorable" phrase sets. Recent work has demonstrated that use of these sets introduces unintended effects when they are not sampled appropriately [11]. As with any lab experiment, participants can be biased towards achieving whatever metric *they* believe the researchers are interested in [15]. It could be argued that if participants assume the metric of interest is accuracy, then they would attempt to use the stimulus phrase not just as a *prompt* at the start of the task, but frequently as a *guide* to ensure low error rates. Kristensson and Vertanen envisaged the potential problem of frequent attention shifts to and from the stimuli [10]. Hiding the stimulus phrase from participants resulted in higher input speed, but also higher error rates, indicating that presence of the stimulus is likely introducing undesired effects.

3 MOTIVATION AND RESEARCH QUESTIONS

Based on the current state of the art, we wondered if we could quantify the impact of permanent stimulus phrase display on participant behaviour. We hypothesised that participants may examine the stimulus phrase often to make constant checks for errors. Just as Kristensson and Vertanen [10] attempted to hide the stimulus phrase to prevent attention issues, we wondered if hiding another part of the user interface could reduce the participants' attention to the stimulus phrase. This could be done by masking the characters of the currently composed word, allowing the user to focus more on the keyboard to complete a word. We therefore formed the following research questions for this paper:

(R1) *How does the persistent presentation of stimulus phrases during a transcription task, affect user attention and behaviour?*, and;

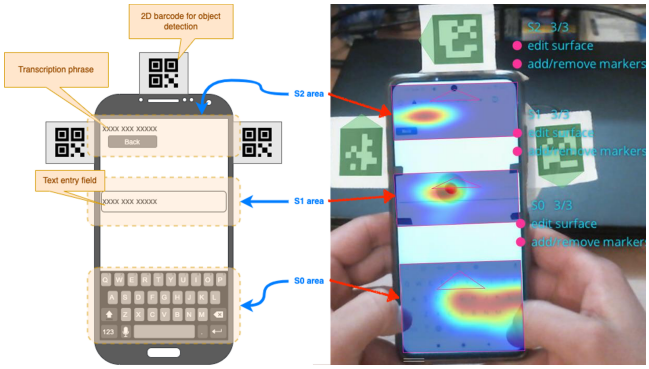
(R2) *Can the masking of input limit user attention to the stimulus phrase area and thus encourage less attention-switching behaviour?*

4 PROCEDURE AND MATERIALS

4.1 Equipment

For our experiment, we used an Android smartphone with Google's GBoard as the text entry method. On the device, we used WebTEM [2] as the experiment software to create a transcription task environment. We worked closely with WebTEM's creator to add two new features for the purposes of the experiment: Firstly, to create a

(a) Eye-tracker used in experiment



(b) Schematics of the areas of interest and fiducial markers setup

Figure 1: Configuration of the experiment environment.

vertical separation between the phrase to be transcribed and the text entry field, placing the latter on the middle of the screen (see Fig. 1b). Further, we implemented the option to mask input characters in a word being composed, replacing them with an asterisk (star) character, and revealing the actual input only after a word separator (e.g., space) had been pressed.

We built the DIY version of the Pupil Labs eyetracker, a low-cost solution fully compatible with Pupil Labs' software for data analysis. The software allows researchers to specify objects of interest within the world view, and to extract gazes and fixations falling within the video frame surfaces covered by these objects. For this purpose, it is required to "mark" objects with a visual fiducial marker and to manually outline one or more areas of interest on the object in the first frame of captured video. We tacked three paper markers on the back of the smartphone's frame, to ensure that at least one of them would be visible to enable tracking in all frames (Fig.1b).

As such, we defined three areas of interest: S0 represents the device keyboard, S1 is the area containing the text entry field and S2 is the stimulus phrase. The areas cover a vertically larger surface than actually occupied by the visual components of interest - this is to allow for calibration drifts and coverage of human foveal vision (the 5° field where visual acuity is at its highest, meaning that a target does not need to be directly on the line of sight in order to be visually perceived in full detail). Gaze and fixation data were exported as comma-separated value files from the Pupil software, and were subsequently processed in Python 3.8 (using Jupyter Notebooks) for the analyses in this paper. We make all raw data and processing code available to other researchers (see Section 8).

4.2 Experiment procedure

4.2.1 Participants. We ran a transcription-based study, using eye-tracking with two cohorts. One consisted entirely of undergraduate students, and one consisted of graduates (professionals), all in the field of Computer Science from the same University department. The first cohort (C1) included 19 participants (10 female), all Computer Science undergraduate students aged between 18-29. All but one users reported using their smartphone to enter text at least "a few times per day" (2) or "very often during the day" (16). The majority (17) self-assessed as fluent or expert English users, though none of the participants spoke English as their native language. We also asked them about use of text entry support features (autocorrect and word completion). The majority of participants reported "always" (6) or "frequently" (9), using these features, while the rest reported using these "occasionally" (2) or "never" (2). For the second cohort (C2), we recruited 25 smartphone users in the age brackets of 18-24 (8 participants) and 25-34 (17), all Computer Science graduates. The majority of users were female (18). The majority (23) self-assessed as fluent or expert English users, though none of the participants spoke English as their native language. All users reported using their smartphone to enter text at least "a few times per day" (11) or "very often during the day" (14). We also asked them about use of text entry support features (autocorrect and word completion). The majority of participants reported "always" (9) or "frequently" (7), using these features, while the rest reported using these "occasionally" (7) or "never" (2).

4.2.2 Experiment protocol. For the experiment, WebTEM was configured to present to each participant with blocks of 10 phrases to be transcribed, picked at random from Kano et al.'s 500 Children's English Phrase set [7]. We employed this phrase set instead of [14, 19], which are more commonly used, in order to reduce complexity for our participants (non-native English speakers). Kano et al.'s set is particularly attractive since it is proven to perform similarly to that of Soukoreff and Mackenzie's [14], with the advantage of being non-location dependent (containing no words with specific American or British spelling) [7]. For each experiment condition (masking - no masking), participants performed three input sessions of 10 phrases, with a short resting break between sessions. Participants started the experiment with conditions presented in a counterbalanced order, to prevent result bias. At the start of the experiment, participants were given time to familiarise themselves with the task, wearing the eye-tracker and using WebTEM to transcribe a few phrases until they indicated they felt comfortable with the setup. No data was recorded in the familiarisation phase. Then, the eye-tracker was calibrated and participants were instructed to begin the experiment, typing as fast and as accurately as possible, with the ability to correct any mistakes if spotted. The eye-tracker was re-calibrated after each block of phrases was completed. Apart from gaze data, we captured the following metrics through WebTEM: Words per Minute (WPM), Error rate, and keystrokes per character.

5 RESULTS

Pupil software allows for the capture of raw gaze and also fixation data. We base our analysis on fixations, which are a more reliable metric of a user's actual attention. Statistical tests were chosen after examination of assumptions for their use.

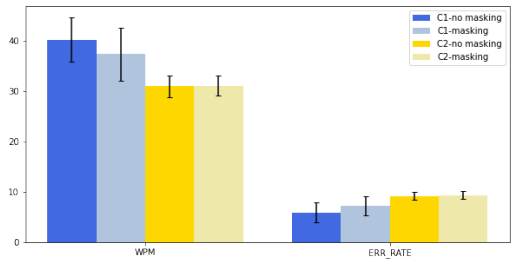


Figure 2: Results of standard text metrics collected by WebTEM. Error bars at 95%c.i.

5.1 Text metrics

Using the data from WebTEM we analysed the basic text entry metrics collected, with results shown in Figure 2. The first observation is that participants in C1 appeared faster and seemed to make fewer mistakes compare to those in C2. Pairwise t-tests in each cohort, showed no statistically significant differences in WPM or Total Error Rate between the masking and no-masking condition. On the other hand, independent sample t-tests for WPM in both masking and no-masking conditions showed that the observed differences between cohorts were statistically significant (WPM-masking $t = 2.480, p = 0.017$; WPM-no masking $t = 3.998, p < 0.001$; TER-masking $t = -2.273, p = 0.028$; TER-no masking $t = -3.325, p = 0.002$). This verifies the initial observation that C1 exhibited better proficiency with the QWERTY keyboard, though it appears that the masking condition did not have an effect on the participants' typing performance, in either cohort.

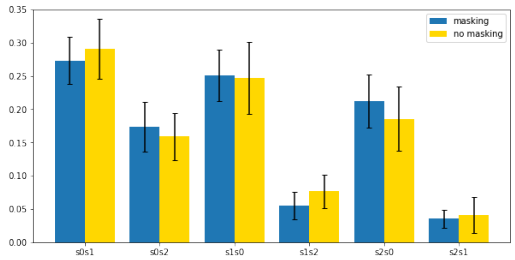
5.2 Division of attention to areas of interest

Our first observation is that a significant proportion of the participants' attention is actually devoted to the stimulus phrase area (S2). As can be seen in Table 1, for C1, approximately one seventh of participant's attention is focused on the stimulus phrase, but this proportion raises significantly for C2, to one quarter. More surprisingly, for C2, this is almost the same as the time devoted to the actual input text. The keyboard area is where participants focus their attention for the majority of the time in C1, and almost half the time for C2. This is a surprising result, since the premise of the transcription task is that presented phrases should be memorable (hence, no need to look up in order to remember the next word to type) and simple (hence, no need for participants to look up complicated spelling). We would have expected that the proportion of attention paid to the stimulus phrase would be very small, compared to the other two areas. In our case, it is clear that participants' behaviour seems to invalidate these assumptions.

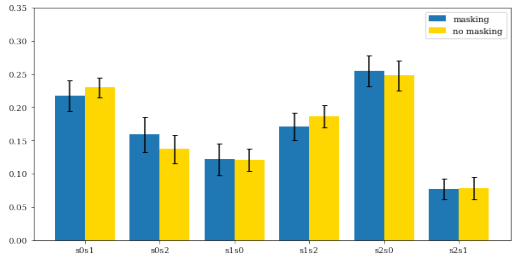
After these results, it is natural to wonder, how does participants' attention shift between these areas of interest? We calculated the number of transitions between areas as follows: by serially traversing through fixations, as ordered by its timestamp, we explored whether it fell on a different surface compared to the fixation immediately preceding it. As such, we calculated the number of transitions between areas for each participant and derived the percentage (see Figure 3). We note that the least frequent transition in both cohorts is from S2-S1 (stimulus to edit area, C1 masking:

Cohort	Area	Masking		No masking	
		$\bar{x}$	σ	$\bar{x}$	σ
C1	S2	13.64%	7.54%	14.09%	7.66%
	S1	19.77%	10.42%	21.87%	13.08%
	S0	66.59%	14.23%	64.04%	14.09%
C2	S2	25.40%	7.60%	25.36%	7.09%
	S1	24.88%	10.54%	25.76%	12.20%
	S0	49.71%	14.25%	48.89%	16.64%

Table 1: Percentage of fixations on the areas of interest.



(a) Cohort 1



(b) Cohort 2

Figure 3: Proportion of fixation transitions between the different areas of interest. Error bars at 95%c.i.

$\bar{x} = 3.55\%, \sigma = 3.20\%$; C1 no masking: $\bar{x} = 4.13\%, \sigma = 3.02\%$; C2 masking: $\bar{x} = 7.78\%, \sigma = 3.91\%$; C2 no masking: $\bar{x} = 7.81\%, \sigma = 4.36\%$;) while the most frequent ones are S0-S1 (keyboard to edit area) in C1 (C1 masking: $\bar{x} = 27.32\%, \sigma = 7.94\%$; C1 no masking:$\bar{x} = 29.03\%, \sigma = 7.94\%$), and S2-S0 (stimulus to keyboard area, C2 masking: $\bar{x} = 25.47\%, \sigma = 5.98\%$; C2 no masking:$24.78\bar{x} = \%, \sigma = 5.86\%$). Pairwise tests show that the differences in the percentage of transition pairs between conditions (masking, no masking) were not statistically significant in either of the cohorts, evidencing again that the masking condition did not seem to have an effect on participant behaviour during text entry.

Figure 3 shows variations within cohorts in the distribution of transitions under each condition. Independent samples t-tests and Mann-Whitney U tests, show the differences all differences between any two areas were statistically significant at the $p < 0.01$ level. Generally, for both conditions, we observe that there was more attention shift between the keyboard and edit areas (S0-S1 and S1-S0) in C1. Inversely, C1 made fewer attention shifts between the edit area and the stimulus phrase (S1-S2 and S2-S1). Additionally,

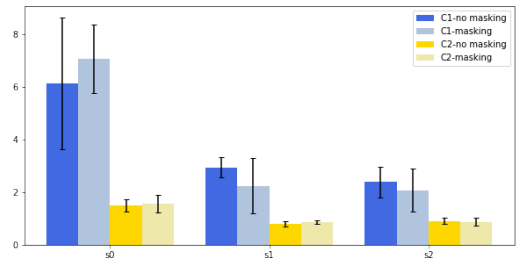


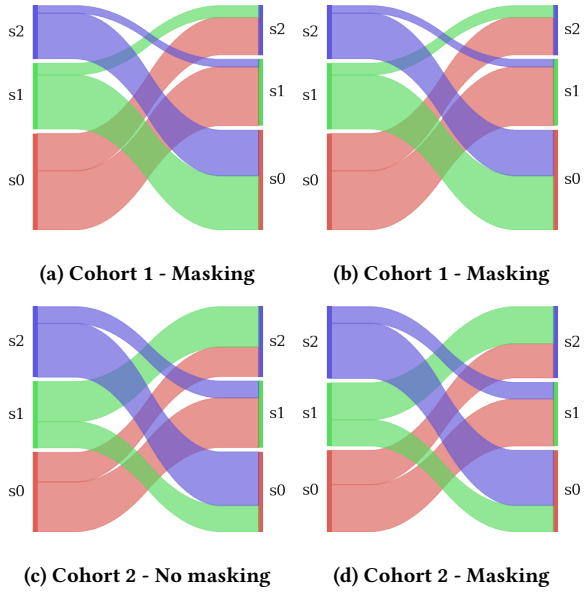
Figure 4: Duration of fixation stints in areas of interest. Error bars at 95%c.i.

C1 made fewer transitions between the stimulus phrase and the keyboard, but only in the Masking condition (S2-S0).

Further, we calculated the length of fixation "stints" in different areas. A stint is defined as the period elapsed between the first and last consecutive fixation in a given area of interest (i.e. before a transition to another area is detected). As shown in Figure 4, C1 made more concentrated efforts, particularly when focussing on the keyboard area, while C2's behaviour seems rather erratic, since the average duration of its fixation stints is very short. Comparing the effect of the masking condition, pairwise Wilcoxon signed rank tests did not find any statistically significant differences in either of the cohorts. Again, it appears that the masking did not have an effect on the participants' behaviour. On the other hand, independent sample Mann-Whitney U tests showed that every comparison was statistically significant, with $p < 0.001$.

Further visualisation of attention flow between areas is shown in Figure 5. Participants in C1, after having focused on the keyboard (S0) are more likely to shift their attention towards the text editing area (S1), compared to the stimulus phrase area (S2) (masking C1:61.12%, C2:57.72%; no masking C1: 64.56%, C2: 62.61%). This is natural error-checking behaviour, since one would want to check what they've typed, after a bit of time spent typing. What is less intuitive though is that a significant percentage of upwards glances from the keyboard, to target the phrase area. The effect is more pronounced in the masking condition (C1: 38.88%, C2: 42.28%) compared to the no masking (C1: 35.44%, C2:37.59%). This can only be interpreted as a participant's wish to see what needs to be typed next, having ostensibly forgotten the rest of the stimulus phrase (remember that the participant does not *need* to memorise the phrase, since it is always available). From the editing area, C1 displays the majority of fixations to transit towards the keyboard (masking: 82.05%, no masking: 76.35%), but this result is reversed for C2, who mostly move from the edit area to the stimulus phrase (masking: 58.43%, no masking: 60.62%). Finally, when the user's attention is at the stimulus phrase area, for both cohorts, the majority of fixations are then steered to the keyboard area (C1 masking: 85.63%, no masking: 81.79%; C2 masking 76.78%, no masking 76.03%).

These results help map out the basic pattern of attention in a transcription task: Users begin by examining the sentence to be transcribed, and then proceed to type some text. After a typing "stint", users will typically check what they've typed in the editing area, to spot any obvious errors. From there, user attention is either diverted to the keyboard area to continue typing (C1), or the



(a) Cohort 1 - Masking **(b) Cohort 1 - Masking**

(c) Cohort 2 - No masking **(d) Cohort 2 - Masking**

Figure 5: Attention flows between areas of interest.

stimulus phrase area (C2), in order to either confirm the spelling of what's been typed, or to affirm the next characters to be typed. This divergence in behaviour in the two groups, perhaps can be partly explained by the lower total error rate in C1. This cohort had a lower total error rate, meaning they could devote fewer mental resources into fixing errors, and therefore perhaps benefited from better working memory recall, thus needing to look at the stimulus less often to remember what they needed to type next. This perhaps implies that observed WPM differences are not due to unfamiliarity with the keyboard, but partly due to the effort required to correct mistakes, and partly due to the need to consult the stimulus phrase more often. A small difference in the total error rate (masking C1-C2 $\Delta\bar{x} = -2.13$; no masking C1-C2 $\Delta\bar{x} = -3.23$) can have a measurable impact of about in terms of typing speed (masking C1-C2 $\Delta\bar{x} = 6.31WPM$; no masking C1-C2 $\Delta\bar{x} = 9.20WPM$). Additionally, the fragmentation of attention exhibited in C2 (shorter attention stints) may have contributed to the overall effect.

Closing this section, a Spearman correlation on WPM, total error rate and the percentage of fixations spent in each area (S0, S1, S2) across all participants, irrespective of cohort, demonstrates that in the no masking condition, text entry speed is statistically significantly correlated only to error rate ($\rho = -0.352, p = 0.20$). Additionally, error rate is also negatively correlated to the percentage of fixations spent in the keyboard area ($\rho = -0.334, p = 0.029$). Further, error rate was negatively with the duration of stints in the keyboard area ($\rho = -0.396, p = 0.009$) and the stimulus phrase ($\rho = -0.318, p = 0.038$). Therefore, without any modifications to the transcription task, participants seem to achieve faster and more accurate input if they focus more on the task of typing, rather than diverting their attention elsewhere.

In the masking condition, entry speed was, surprisingly, not statistically significantly correlated to error rate. Instead, it was negatively correlated to the amount of percentage of fixations spent on the stimulus phrase ($\rho = -0.367, p = 0.016$) while error rate was again only negatively correlated to the percentage of fixations spent in the keyboard area ($\rho = -0.330, p = 0.031$). The duration of fixation stints in the keyboard area and editing area were positively correlated with entry speed (S0:$\rho = 0.322, p = 0.035$; $S1 : \rho = 0.470, p = 0.001$). Error rate was also negatively correlated with the duration of fixation stints in all areas (S0:$\rho = -0.439, p = 0.003$; $S1 : \rho = -0.320, p = 0.036$; $S2 : \rho = 0.360, p = 0.018$). Our modification to the transcription task required participants to focus on all areas of the smartphone, in order to be more successful.

6 DISCUSSION

In this paper, we implemented a mechanism to discourage frequent error-checking behaviour, but did not achieve the envisioned effect. This could be due to the short word length and simplicity in the phrase sets, which would make the masking less relevant than if participants had to enter unfamiliar or complicated text. The results might be more pronounced if the masking was applied to longer sequences of text (e.g. 2-3 words).

More importantly we demonstrate that participant fixations on the stimulus phrase represent a significant proportion (15-25%) of total fixations during the transcription task. Vertanen and Kristensson warned of the potential dangers of excessive user attention to the persistent display of the stimulus phrase [10, 18]. Our work substantiates this warning and, for the first time, quantifies the cost of visual attention to the stimulus phrase. We find that this attention cost is incurred differently across two cohorts of partipants, meaning that individual users, or groups of users may be differently impacted during a transcription task. We used a number of participants in two cohorts (19+25) which was in line with relevant literature (e.g. 16 in [16], 21 in [8], 33 in [6]). The differences that were discovered, highlight that text entry experiment results are subject to variation, even between seemingly similar samples, and therefore caution must be made in claims of generalisability. Studies with participants from various backgrounds would be needed for stronger generalisability. As non-native speakers, our participants might have also felt less secure about their language skills during entry, and thus checked the stimulus phrases more often.

Given the aforementioned limitations, we believe that for the current state of text entry research, our findings have the following implications. An important sub-component of entry speed is inter-key interval (IKI), and where long IKIs are displayed (e.g. at the end of words), they can now be explained as owed, up to 50%, to time spent examining the stimulus phrase, either to check spelling, or to remember what needs to be typed next. A corollary is that we can possibly derive a better metric for text entry speed in transcription tasks, by excluding participant time spent on checking the stimulus phrase. The observation of low uncorrected error rates in transcription tasks can now also be explained. Since the participants' attention is so focused on the stimulus phrase, it is unsurprising that very few, if any, spelling mistakes are left uncorrected. This enhances the argument in [3] that only the Total Error Rate metric should be used in text entry studies, which combines both the effort incurred in detecting and correcting mistakes, with those left unattended. As per the previous points, we need to wonder

whether the transcription task needs metrics of its own, which may not be applicable to other types of evaluation method (particularly in-the-wild). We might need to carefully consider how to improve the transcription task, so that its external validity is enhanced.

7 CONCLUSIONS

We find that participant attention to stimulus phrases during transcription experiments is high, accounting for up to 25% of total attention during task execution, and 50% of the attention devoted outside the keyboard area. This finding raises several questions on the continued use of the transcription task as the de-facto evaluation method for mobile text entry research. It appears that stronger participant focus on the task of composing text on the keyboard, leads to better performance in terms of both speed and accuracy. Contrary to the demand of being as accurate as possible in transcribing a given phrase, a better assessment of participants' actual performance might be to allow them to enter text as they see fit. Even if a stimulus phrase was used as a prompt with the instruction to copy it, there is no reason why it should be copied *exactly* as given. Error rates might be calculated against the stimulus, but only for those parts that the participant indeed intended to copy precisely. Omitted, altered or added words, should probably not count towards error metrics, as long as participants entered that text with good accuracy. Such changes to how we structure and evaluate transcription tasks, might mitigate issues related to anxiety to perform, or short-term memory limits, and thus provide more reliable results for text entry research. We might thus wonder whether the community should devote effort to either significantly improve, or altogether replace the transcription task. In [20] the authors argued that the *Composition* task should be a complement, and not a direct replacement for the transcription task. Since the largest advantage of the transcription task is that user input can be compared for accuracy against a known piece of text, and given that this metric can be "gamed" by the presence of the stimulus phrase, is there a reason why we should want to continue with transcription tasks as the main evaluation method? Perhaps it's time to move on.

8 DATA AND CODE AVAILABILITY

Data and processing code (Python) available at https://github.com/komis1/dontlookup.

ACKNOWLEDGMENTS

We are grateful to A.S. Arif for generously accepting to accommodate the necessary modifications to WebTEM [2].

REFERENCES

[1] Ahmed Sabbir Arif, Sunjun Kim, Wolfgang Stuerzlinger, Geehyuk Lee, and Ali Mazalek. 2016. Evaluation of a Smart-Restorable Backspace Technique to Facilitate Text Entry Error Correction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5151–5162. https://doi.org/10.1145/2858036.2858407

[2] Ahmed Sabbir Arif and Ali Mazalek. 2016. WebTEM: A Web Application to Record Text Entry Metrics. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces (ISS '16)*. Association for Computing Machinery, New York, NY, USA, 415–420. https://doi.org/10.1145/2992154.2996791

[3] A. S. Arif and W. Stuerzlinger. 2009. Analysis of Text Entry Performance Metrics. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*. 100–105. https://doi.org/10.1109/TIC-STH.2009.5444533

[4] Nikola Banovic, Ticha Sethapakdi, Yasasvi Hari, Anind K. Dey, and Jennifer Mankoff. 2019. The Limits of Expert Text Entry Speed on Mobile Keyboards with Autocorrect. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3338286.3340126

[5] Leah Findlater, Joan Zhang, Jon E. Froehlich, and Karyn Moffatt. 2017. Differences in Crowdsourced vs. Lab-based Mobile and Desktop Input Performance Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, Denver, Colorado, USA, 6813–6824. https://doi.org/10.1145/3025453.3025820

[6] Xinhui Jiang, Yang Li, Jussi P.P. Jokinen, Viet Ba Hirvola, Antti Oulasvirta, and Xiangshi Ren. 2020. How We Type: Eye and Finger Movement Strategies in Mobile Typing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376711

[7] Akiyo Kano, Janet C Read, and Alan Dix. 2006. Children's Phrase Set for Text Input Method Evaluations. In *Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles (NordiCHI '06)*. Association for Computing Machinery, New York, NY, USA, 449–452. https://doi.org/10.1145/1182475.1182534

[8] Huhn Kim, Seungyoun Yi, and So-Yeon Yoon. 2019. Exploring Touch Feedback Display of Virtual Keyboards for Reduced Eye Movements. *Displays* 56 (Jan. 2019), 38–48. https://doi.org/10.1016/j.displa.2018.11.004

[9] Andreas Komninos, Emma Nicol, and Mark Dunlop. 2020. Investigating Error Injection to Enhance the Effectiveness of Mobile Text Entry Studies of Error Behaviour. *arXiv:2003.06318 [cs]* (March 2020). arXiv:2003.06318 [cs] http://arxiv.org/abs/2003.06318

[10] Per Ola Kristensson and Keith Vertanen. 2012. Performance Comparisons of Phrase Sets and Presentation Styles for Text Entry Evaluations. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces (IUI '12)*. Association for Computing Machinery, New York, NY, USA, 29–32. https://doi.org/10.1145/2166966.2166972

[11] Luis A. Leiva and Germán Sanchis-Trilles. 2014. Representatively Memorable: Sampling the Right Phrase Set to Get the Text Entry Experiment Right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, Toronto, Ontario, Canada, 1709–1712. https://doi.org/10.1145/2556288.2557024

[12] Yang Li, Sayan Sarcar, Sunjun Kim, and Xiangshi Ren. 2020. Swap: A Replacement-based Text Revision Technique for Mobile Devices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376217

[13] I. Scott MacKenzie and R. William Soukoreff. 2002. Text Entry for Mobile Computing: Models and Methods,Theory and Practice. *Human–Computer Interaction* 17, 2-3 (Sept. 2002), 147–198. https://doi.org/10.1080/07370024.2002.9667313

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

[15] Jim McCambridge, Marijn de Bruin, and John Witton. 2012. The Effects of Demand Characteristics on Research Participant Behaviours in Non-Laboratory Settings: A Systematic Review. *PLOS ONE* 7, 6 (June 2012), e39116. https://doi.org/10.1371/journal.pone.0039116

[16] Felix Putze, Tilman Ihrig, Tanja Schultz, and Wolfgang Stuerzlinger. 2020. Platform for Studying Self-Repairing Auto-Corrections in Mobile Text Entry Based on Brain Activity, Gaze, and Context. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376815

[17] Shyam Reyal, Shumin Zhai, and Per Ola Kristensson. 2015. Performance and User Experience of Touchscreen and Gesture Keyboards in a Lab Setting and in the Wild. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 679–688. https://doi.org/10.1145/2702123.2702597

[18] 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 and Services (MobileHCI '11)*. Association for Computing Machinery, Stockholm, Sweden, 295–298. https://doi.org/10.1145/2037373.2037418

[19] 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 and Services (MobileHCI '11)*. Association for Computing Machinery, New York, NY, USA, 295–298. https://doi.org/10.1145/2037373.2037418

[20] Keith Vertanen and Per Ola Kristensson. 2014. Complementing Text Entry Evaluations with a Composition Task. *ACM Trans. Comput.-Hum. Interact.* 21, 2 (Feb. 2014), 8:1–8:33. https://doi.org/10.1145/2555691