How do People Type on Mobile Devices? Observations from a Study with 37,000 Volunteers

Kseniia Palin
Aalto University, Finland

Anna Maria Feit
ETH Zurich, Switzerland

Sunjun Kim
Aalto University, Finland

Per Ola Kristensson
University of Cambridge, UK

Antti Oulasvirta
Aalto University, Finland

ABSTRACT

This paper presents a large-scale dataset on mobile text entry collected via a web-based transcription task performed by 37,370 volunteers. The average typing speed was 36.2 WPM with 2.3% uncorrected errors. The scale of the data enables powerful statistical analyses on the correlation between typing performance and various factors, such as demographics, finger usage, and use of intelligent text entry techniques. We report effects of age and finger usage on performance that correspond to previous studies. We also find evidence of relationships between performance and use of intelligent text entry techniques: auto-correct usage correlates positively with entry rates, whereas word prediction usage has a negative correlation. To aid further work on modeling, machine learning and design improvements in mobile text entry, we make the code and dataset openly available.

CCS CONCEPTS

• **Human-centered computing** → Empirical studies in HCI.

KEYWORDS

Mobile text entry; word prediction; auto-correct

ACM Reference Format:
Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How do People Type on Mobile Devices? Observations from a Study with 37,000 Volunteers. In *21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19), October 1–4, 2019, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3338286.3340120

MobileHCI '19, October 1–4, 2019, Taipei, Taiwan

ACM ISBN 978-1-4503-6825-4/19/10…$15.00
https://doi.org/10.1145/3338286.3340120

1 INTRODUCTION

This paper contributes to efforts in understanding typing performance with mobile devices, a central topic in recent HCI research (e.g. [3, 5, 7, 29, 33, 35]). Mobile devices are extensively used for text input, in activities such as email, internet browsing, texting, and social media [9]. However, mobile typing is generally slower than typing on physical keyboards [35]. Existent literature attributes this to a number of factors (see *Related Work*), including the use of virtual instead of physical buttons, the use of fewer number of fingers, and the absence of training regimes like the ten-finger touch typing system. At the same time, a large number of intelligent text entry techniques exist, the effect of which is poorly charted beyond prototype evaluations.

We present a new large-scale dataset and first observations of correlates of typing performance. To improve text entry techniques, it is important to understand their effects beyond controlled laboratory studies. While most studies in HCI have involved a relatively low number of participants [8], and often focused on prototype evaluation, we report here results from a large-scale dataset of over 37,370 volunteers. Large-scale analyses of mobile interaction are relatively rare and mostly undertaken by commercial organizations that may keep the datasets proprietary. Such analyses can contribute to more comprehensive statistical analyses of a larger number of interacting variables, and serve as training data for machine learning models. However, self-selection bias is a real threat to generalizability of results in online studies with volunteers or paid workers. To this end, we report on participant demographics and perform stratified subsampling that allows for partial bias mitigation and better estimation of population distribution.

In this work, we first present the data collection method and describe the dataset. We then report on distributions of common metrics of typing performance, including words per minute (WPM), error rate, and keystrokes per character (KSPC). To better understand mobile typing behavior, we present observations on the impact of demographic factors and typing styles on performance. In particular, we report on previously underexamined relation with intelligent text entry techniques (ITE), such as autocompletion, gestural

text entry, and word prediction. Our key findings are: (1) when compared to small-scale studies of mobile typing, performance in practice seems to be higher than previously reported; ca. 36.2 WPM on average in our study; (2) differences in age, experience with English, and typing style impact performance, whereas prior training of touch typing for desktop keyboards does not; and (3) intelligent text entry methods have a varied effect: autocorrection appears to be positively associated with performance while word prediction negatively.

2 RELATED WORK

A large body of research has emerged seeking to understand factors affecting performance and to find techniques to assist typing on mobile devices. In this section, we briefly review some main earlier results.

Factors Affecting Mobile Typing Performance

Many studies of typing performance on virtual keyboards have been conducted to evaluate a new input technique or to collect training data for models. Some studies exist that investigate typing behavior or background factors impacting typing performance. Generally, it is known that typing with one-finger is slower than using two thumbs. Azenkot and Zhai reported speeds of 50.03, 36.34, and 33.78 WPM when entering text with two thumbs, one thumb or the index finger, respectively [3]. The superiority of two-thumb input is attributed to frequent switching between the sides of the display, which allows for preparatory movements that decrease inter-key intervals [6, 26, 28, 33].

Errors in mobile typing are costly in comparison to physical keyboards. The lack of tactile feedback makes it hard to recognize pointing errors even when focusing on the virtual keyboard. Thus, users need to shift attention from the keyboard to the input field to detect the mistakes and back to the keyboard to correct them. The more often one looks at the input field, the more quickly one can detect an error. However, typing will be slower as attention is needed to guide fingers on the display and editing the input field is often cumbersome. Recent work found that this may cause adjustments in a speed-accuracy trade-off. Users may for example slow down to minimize the risk of errors [5]. Cognitive and motor impairments, such as dyslexia, tremor, or memory dysfunction, and various effects of aging, can have a detrimental effect on typing performance. Users adjust their sensorimotor strategies to find a suitable compromise between speed and accuracy and reliance on intelligent text entry techniques such as the word prediction list [32].

Two recent studies report mobile typing behavior in situ. Buschek et al. [7] conducted a study of 30 people using a customized keyboard over three weeks, which sampled and logged the typed text in a privacy-preserving mode. They reported an average typing speed of 32.1 WPM. 74% of typing was done using two thumbs, only 12.7% with the right thumb and less than 3% for all other hand postures. Twenty-seven participants used the word prediction feature, on average for about 1.6% of the entered words. Sixteen used the autocorrection features. Only 0.63% of keystrokes were performed in the landscape mode. Komninos et al. [19] conducted a field study (N = 12) using a customized keyboard over 28 days. They reported on average ca. 34 keystrokes per typing session, with 1.98 uncorrected words. If mistakes were noticed, most were corrected by using 1-5 backspace keystrokes. However, the generalisability of these observations is limited by the size and composition of the samples (e.g., low number of participants sampled from technical fields of a single country [7]).

Intelligent Text Entry Methods

ITE methods use statistical language models to exploit the redundancies inherent in natural languages to improve text entry. Such improvements can be channeled to the user in various ways. For example, an ITE method can autocorrect previous typing, complete on-going input or predict the next word for the user (see [21] for a brief review).

Numerous ITE methods have been presented in the literature and are implemented in commercial keyboards Many aim at improving input accuracy, and thus speed, for example by correcting touch points [15, 16], resizing key targets [14, 15], creating personalized touch models [40, 43], taking into account individual hand postures and finger usage [3, 13, 27, 43], or by adapting to walking speed [27]. Statistical decoding to auto-correct users' typing has been demonstrated to be quite powerful, such as in the context of smart watch typing [39].

However, the efficacy of word prediction is unclear for mainstream mobile text entry, For example, the user has to switch attention from the keyboard and the typed text to the word prediction list. Usefulness is therefore determined in a complex interplay of many factors, including the efficiency of the used text input method, the experience of the user, the accuracy of the prediction and other factors. Accordingly, results reported in the literature have been mixed [1, 18, 34]. In particular, for typing on mobile keyboards, a recent study showed that the use of word prediction methods can be detrimental to performance [29].

Gesture keyboard entry (originally called SHARK or Shape writing) [20, 23, 44], where users continuously draw from one letter of a word to another, permits the use of gestures that are argued to evolve with repetition into fast-to-execute open-loop motor programs. When assessed outside the lab, people performed almost 10 WPM faster after practice than

using tapping-based input [31]. While gesture keyboard entry is a mainstream text input method, some research indicates that it does not experience frequent use in practice [7].

Many of these techniques have been tested in controlled laboratory evaluations with small sample sizes. The generalizability of these benefits to a broader population is less understood. Our dataset presented gives insights into the use of ITE methods across a broad population. Moreover, it can serve as training data to improve these methods and to develop new techniques.

Methods for Studying Mobile Typing

Our methodology closely follows prior work on large-scale online studies of desktop typing [11]. To assess a user's typing speed, we use transcription typing, a common task to study motor performance that excludes cognitive aspects related to the process of text generation. Using an online typing test allows us to reach a larger and more diverse set of people [30] than would be possible in a lab study. In comparison to desktop settings, mobile typing studies have also been frequently conducted outside the lab (e.g. [7, 16, 19, 31]). This often allows observation of more realistic behavior of users and thus yields different insights in comparison to lab studies (e.g. as discussed in [31]). Using an online platform allows us to reach a much larger number of participants in-situ. Similar approaches have been used for example to collect large amount of training data for creating touch models [16]. In comparison to most prior work discussed above, we do not require users to install a dedicated app. Instead, we offer an online typing test for users to assess their typing performance using any keyboard they are comfortable with and any ITE method they are used to. This also allows us to reach an even larger and more diverse sample of participants than previous in-situ studies.

3 DATA COLLECTION

Data were collected in a web-based transcription task hosted on a university server. A web-based method, as opposed to a laboratory- or app-based data collection, permits a larger sample and broader coverage of different mobile devices, but comes with the caveats of self-selection and compromised logging accuracy (see below). Still, the typing test setting imposes a more controlled environment than an in-the-wild study. Our test supports the main mobile operating systems and browsers and was available globally on the Internet in a collaboration with a Web company offering typing testing and training. In the design of the software, we directly built on work by Dhakal et al. [11] who studied transcription typing on physical keyboards: (1) we used the same phrase set representative of the English language; (2) we updated performance feedback only after users committed a phrase; and (3) we chose well-understood performance metrics covering speed and errors. However, we needed to adapt the software to support mobile devices, making it responsive to different screen sizes, changing the logging and updating the database structure. Also, we added questions regarding the keyboards used, people's typing posture, and the use of ITE methods. In our analysis, we perform subsampling to mitigate the self-selection bias.

Table 1: Summary of demographics and typing-related background factors in the full sample and the U.S. subsample after pre-processing. SD shown in brackets.

	Full sample		U.S. subsample	
Factor	Result	Remark	Result	Remark
Gender	65/31 % f/m	4% n/a	51/49 % f/m	
Age	24.1 (8.8)	75% 28 yrs.	25.7 (12.3)	75% 32 yrs.
Countries	163	47% U.S.	1	100% U.S.
En native speakers	68%		88.2%	
Typing course	31.4%	Desktop	40.1%	Desktop
H/day typing	6.5 (6.2)	on mobile	5.6 (5.9)	on mobile
Detected OS	51/49 % Android / iOS		53/47% Android / iOS	

Participants

Our participants volunteered via a public website[1] for training and testing of typing skills. HTML requests to the site that originated from devices detected as mobile (screen width $< 800\ px$), were redirected to our test. The data were collected between September 2018 and January 2019.

Table 1 summarizes the demographic background of the 37,370 voluntary participants left in the database after preprocessing (see below). Similar to the general user-base of the company hosting the webpage (see [11]), the test was completed by more females than males, the majority of which came from the U.S. and were mostly experienced in typing in English (native - 56%, always - 21%, usually - 12%). The majority reported entering text using two thumbs (74%) and using the QWERTY layout (87%). Most did not use third-party keyboard apps (79%). Android and iOS devices were used almost equal to Mobile Safari (43%) and Chrome Mobile (38%) browsers used most often.

U.S. subsample (N = 1475). In the rest of the paper, we report comparative statistics from a stratified subsample that better matches the general population of the U.S., the best-represented country in our sample. We randomly selected U.S. participants to match the distributions of gender [10], age groups [10], and mobile operating systems [17], resulting in a subsample of 1475 participants. See Table 1 for details.

[1] https://www.typingtest.com/

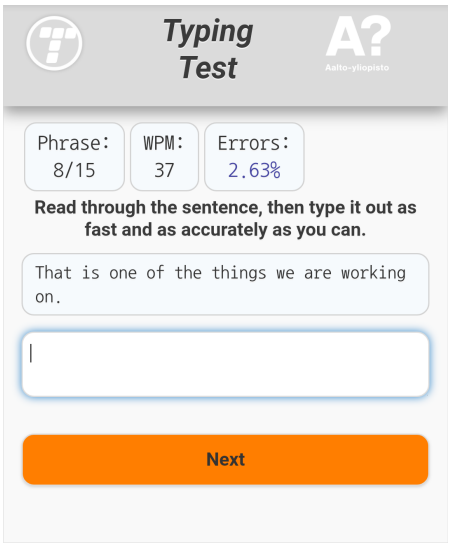

Figure 1: The web-based transcription task. One sentence was presented at a time with the progress shown at the top.

Task and Procedure

We followed the same procedure as Dhakal et al. [11]. The task was to transcribe 15 English sentences, shown one after another. Participants were shown instructions requesting they first read and memorize the sentence, then type it as quickly and accurately as possible. Breaks could be taken between the sentences. After acknowledging that they had read the instructions and giving their consent for data collection, the first sentence was displayed. Upon pressing Next or the Enter key, the user's progress, their speed and error rate were updated and the next sentence was shown. The sentence was visible at all times. The user interface is shown in Figure 1. When all sentences had been transcribed, participants were asked to fill in a questionnaire before they were shown their final result. In addition to the questions related to demographics and typing experience asked by Dhakal et al. [11], we also asked for their typing posture (1- or 2-hand, index finger(s), thumb(s) or other) the keyboard app and layout they used, and whether they used autocorrection, word prediction, or gesture typing (see below). Then, performance results were shown as a histogram over all participants with details on the fastest/slowest and most error-prone sentences (see [11] for details). Finally, participants were offered to transcribe more sentences to improve the performance assessment, which we did not include in the following analysis.

Material

We used the same sentences as [11], drawn randomly from a set of 1,525 sentences, composed of the Enron mobile email corpus (memorable set from [37], 400 sentences) and English Gigaword Newswire corpus. The former one is representative of the language people use when typing on mobile devices but too small to be used alone. The latter one complements the set with more diverse sentences with a higher Out-Of-Vocabulary (OOV) rate (0.8% versus 2.2% [37]). Mobile text entry can exhibit much higher OOV rates (e.g. >20% on Twitter [4]) with respect to a general text corpus. However, modern ITE methods adapt to users' vocabularies. We thus assume that the low OOV rates of our sentences are representative of mobile text input in practice.

Implementation

We implemented the front-end of our typing test using HTML, CSS, and JavaScript. The back-end was implemented in Scala via the Play framework, using a MySQL database for storing the timestamp, key characteristics, and state of the input field at every key press, as well as meta-data of the participant and session. The data were stored on the same university owned server where the application was hosted on.

Limitations of web-based logging. In contrast to typing on Desktop keyboards [11] which redirect raw device-level events to the input field, the access privileges of web applications are limited on most mobile devices. Similar to other online transcription tests [2], our browser-side logging has the following limitations: (1) the keycode is reported as `undefined` for some Android devices[2]; (2) for many devices touch-down and -up events are generated together at the moment of touch-up, resulting in a keystroke duration of <10 ms; (3) in the case of multi-touch / rollover [11], the events are not transmitted correctly: the key-up event of the first keystroke is dispatched as soon as the second finger touches the screen, although the first key is still pressed down. As a result, the keycode of pressed key was often not available and the accuracy of timestamps was poor. To ensure a consistent analysis, we did not analyze metrics related to the timing of individual keystrokes, such as inter-key intervals, keystroke durations, and rollover ratio [11]. Similar to prior work (e.g. the WebTEM application [2]), we inferred the pressed key from changes in the text on input field.

4 DATA PREPROCESSING AND ANALYSIS

The collected dataset was preprocessed to remove incomplete, inaccurate, or corrupted items. We included only participants who finished 15 sentences and completed the questionnaire. This only included 19% of the over 260,000 people that started the typing test. Such high dropout rates are common in online studies [11, 30]. Of these, we conservatively excluded about 25% of participants who did not use a mobile device, who reported to be younger than than five or older than 61 years (more than 2 SD away from mean age), whose

[2]test e.g. at https://w3c.github.io/uievents/tools/key-event-viewer.html

average typing speed was over 200 WPM, who left more than 25% uncorrected errors, or who took long breaks within a sentence (inter-key interval >5s). This yielded a dataset of 37,370 participants typing 15 sentences each.

Analyzed Metrics

We followed the de facto standard definition of performance metrics in text entry research [41], where some of the following metrics were already computed during runtime. Further analysis was conducted using these measures, computed per sentence and then averaged for each user:

Words per minute (WPM). Computed as the length of the input (one word defined as five characters) divided by the time between the first and the last keystroke.

Keystrokes per character (KSPC). The number of input events (including non-scribed key presses) divided by the number of characters in the produced string.

Uncorrected error rate (ER). Calculated as the Levenshtein edit distance [24] between the presented and transcribed string, divided by the larger size of the two strings. The uncorrected errors were further classified into *insertion*, *omission*, and *substitution* errors as suggested by MacKenzie and Soukoreff [25] using the TextTest tool [42].

Backspaces (BSP). The average number of backspace presses per sentence.

Recognition of ITE from Logs

As described above, web-based logging is limited in the information it receives about each keystroke. As a result, we could not reliably identify the ITE methods from the dispatched events. Instead, we had to rely on the changes in the input field for inferring the use of ITE methods. Therefore, we developed a simple but effective heuristics-based recognition scheme. It compares the state of the input field before and after an input event. Therefore, it uses the last character of the input field, the length of the text and the Levenshtein edit distance, which captures the amount of change in the input field. Each input event is characterized as one of four events:

None: is a "normal" keystroke where no ITE method was used. We recognize this event in the case where only a single character was inserted.

Autocorrection (A):. is the event where the keyboard automatically changes the word after the user finishes it (e.g. by pressing space). We recognize autocorrection if the previous input was a normal keystroke and then multiple characters were changed while the length of the text remained about the same.

Table 2: Confusion matrix of ITE recognition. A = Autocorrection; P = Prediction; G = Gesture.

		Recognized as			
		A	P	G	none
True	A	**32**	1	0	1
	P	3	**71**	0	13
	G	1	7	**425**	27
	none	7	23	21	**7022**

Prediction (P):. is the event where a user finishes the currently typed word by selecting it from a word prediction list. We recognize prediction if the previous input was a normal keystroke, multiple characters were changed, and the length of text increased by more than two characters.

Gesture (G):. is the event where the user continuously draws from one letter to another to input a full word. We recognize a gesture if a whole word is inserted after a space character input or after a gesture.

The exact algorithm used to recognize each input event is available at https://userinterfaces.aalto.fi/typing37k. Note, that the definition of these events corresponds to the interaction of the user: in the case of autocorrection, the user does not perform any additional action, while prediction requires the user to actively shift their attention to the word prediction list and select the right word, and the use of gestures requires them to change their input actions from tapping to swiping. From an algorithmic point of view, these events might be entangled. For example, the keyboard might apply autocorrection to the detection of a gesture, in which case only "Gesture" is recognized.

Empirical validation: To validate our ITE recognition, we collected ground truth data from fifteen volunteers who did the typing test in our lab using their mobile device. Each ITE method was used by at least five participants. We externally recorded the device's screen and manually labeled the ITE input they used for each keypress.

Like this, we collected 7,654 manually labeled input events: 34 autocorrections, 87 predictions, 460 gestures and 7,073 none-ITE inputs. We labeled the events with our recognition algorithm; Table 2 shows the confusion matrix. Overall, the algorithm recognized 90.9% of ITE events correctly (=9.1% false-negative rate), with a low false-positive rate of only 0.7% (none-ITE events recognized as any of the ITE methods).

5 RESULTS

We report on indicators of typing performance and analyze how demographic factors, typing behavior, and use of ITE are associated with performance. Since most of our data are not normally distributed, we used the Mann-Whitney U test

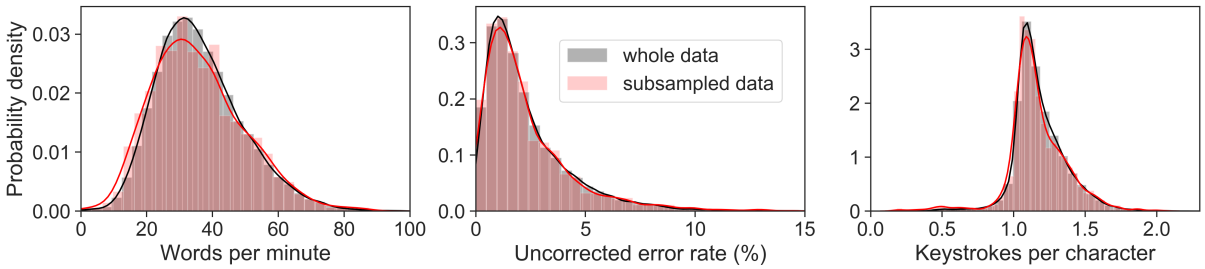

Figure 2: Histogram and density estimate of WPM (left), uncorrected error rate (middle), and KSPC (right) over all data and the U.S. subsample.

to test differences between distributions. In the case of more than two groups, we first used a Kruskal-Wallis test; if a significant difference was found we performed follow up pairwise comparisons using the Mann-Whitney U test with Holm-Bonferroni correction. Effect sizes are reported using Cohen's d.

Performance Measures

An overview of the performance measures overall data in comparison to the U.S. subsample is given in Table 3.

Words per minute. On average, participants typed at 36.17 WPM ($SD = 13.22$) with 75% of participants having a performance below 43.98 WPM. The fastest typists reached over 80 WPM. Figure 2 shows the distribution over all participants. It has skewness of 0.72 and kurtosis of 1.13. The average WPM of participants in the U.S. subsample is similar, 35.99 WPM ($SD = 14.15$). The distribution shown in Figure 2 is slightly different, with a skewness of 1.08 and a kurtosis of 4.14.

Uncorrected error rate. On average, participants left 2.34% (SD=2.08) of errors uncorrected; 75% of participants left less than 3.07%. The skewness of the distribution shown in Figure 2 is 2.54, kurtosis is 12.12. The distribution of the U.S. subsample is similar to an average uncorrected error rate of 2.25% (SD=2.04). The skewness of the distribution is 3.02, kurtosis is 18.65. The uncorrected error consisted of 11.1% insertion errors, 55.6% substitution errors, and 33.3% omission errors. Substitution was the most salient error type, which is in line with a study of text entry on physical keyboards [11].

Keystrokes per character. The average KSPC value for participants is 1.18 ($SD = 0.18$), similar to that of the U.S. subsample ($M = 1.17$, $SD = 0,2$). In both, 75% of participants made less than 1.28 keystrokes per character. Figure 2 shows a similar distribution for both groups.

Backspaces. On average, participants performed 1.89 backspaces per entered sentence, with a large standard deviation of 1.96. Participants in the U.S. subsample performed fewer corrections, a statistically significant difference, but with a small effect size.

Discussion. Typing performance in our sample is relatively high in comparison to prior smaller-sample studies that required participants to use a dedicated app and keyboard. They reported input rates of 32 WPM [7] and 31 WPM [31].

The large sample allows us to make a statistically reliable comparison of typing on mobile soft keyboards versus physical desktop keyboards, which were studied with the same method by Dhakal et al. [11]. Average performance is about 15 WPM slower in mobile typing. Participants left more errors uncorrected on mobile devices (2.34% versus 1.17% [11]). Accordingly, the amount of backspacing is also lower (1.89 versus 2.29 on average). A possible explanation is the higher interaction cost of correcting mistakes on mobile devices and the limited text editing methods (see [5] for a discussion). Nevertheless, KSPC is remarkably similar ($M = 1.17$, $SD = 0.09$ in [11]) with only the standard deviation being smaller compared with desktop keyboard entry potentially due to the varying use of intelligent text entry methods on mobile devices (see below).

Note that we did not report corrected errors here. There is no standard metric that allows to include ITE methods because ITE input breaks the assumption of keystroke level analysis. We call for future work to develop metrics that take this into account, as this is beyond the scope of this paper. However, BSP is a related metric to corrected error.

Table 3: Typing performance of the participants, for the full sample and in the U.S. subsample.

	Overall		U.S. subsample		Statistics		
	$\overline{X}$	σ	$\overline{X}$	σ	p	d	Sign.
WPM	36.17	(13.22)	35.99	(14.15)	.045	.04	*
ER	2.34	(2.08)	2.25	(2.04)	.716	–	–
KSPC	1.18	(0.18)	1.17	(0.20)	.061	–	–
BSP	1.89	(1.96)	1.70	(1.84)	.002	.09	**

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, d: Cohen's d value

Table 4: Overview of WPM for different demographic factors overall and for the U.S. subsample. Significance is indicated for differences within a demographic factor. For factors with more than two groups, detailed statistics of pairwise comparisons are given in the supplementary material.

	WPM all			WPM U.S.		
	$\overline{X}$	σ	Stat.	$\overline{X}$	σ	Stat.
Gender			**			***
female	36.0	(12.5)	p=.002	34.6	(12.3)	p<.001
male	36.1	(14.4)	d=.003	37.4	(15.7)	d=.22
Age			***			***
10-19	39.6	(14.3)	p<.001	38.0	(14.8)	p<.001
20-29	36.5	(12.6)		39.0	(13.3)	
30-39	32.2	(10.8)		34.3	(13.8)	
40-49	28.9	(9.2)		27.3	(9.0)	
50-59	26.3	(9.9)		24.8	(9.1)	
Language use			***			*
native	37.8	(13.6)	p<.001	36.5	(14.5)	p=.03
always	35.9	(13.1)		35.7	(12.7)	
usually	34.5	(11.8)		33.5	(14.1)	
sometimes	30.4	(10.5)		28.8	(12.2)	
rarely	29.6	(11.2)		20.5	(9.4)	
never	25.6	(12.4)		26.4	(1.2)	
Training			***			*
no	36.4	(13.1)	p<.001	36.0	(14.1)	p=.026
yes	35.7	(13.4)	d=.05	35.9	(14.2)	d=.002
Fingers			***			***
2	37.7	(13.2)	p<.001	37.9	(13.8)	p<.001
1	29.2	(10.7)	d = .66	28.6	(12.9)	d = .65
Posture			***			***
both, thumbs	38.0	(13.1)	p<.001	38.2	(13.7)	p<.001
both, index	32.6	(12.7)		32.7	(13.3)	
right, thumb	30.2	(10.5)		30.1	(10.7)	
right, index	26.7	(9.7)		25.4	(9.8)	
left, thumb	30.8	(12.4)		28.9	(11.1)	
left, index	25.0	(11.4)		21.8	(4.2)	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, d: Cohen's d value

Demographic Factors

We analyze the differences in WPM between different population groups categorized by gender, age, use of language, typing training, and finger usage. Table 4 summarizes the results. Details of all the statistical tests for all pair-wise comparisons (Holm-Bonferroni corrected) are available on our project page userinterfaces.aalto.fi/typing37k.

Gender: Average performance of men and women was similar. They both typed at about 36 WPM with only the SD of WPM being smaller for female typists. Note that this analysis excludes 4.6% of participants who did not report their gender.

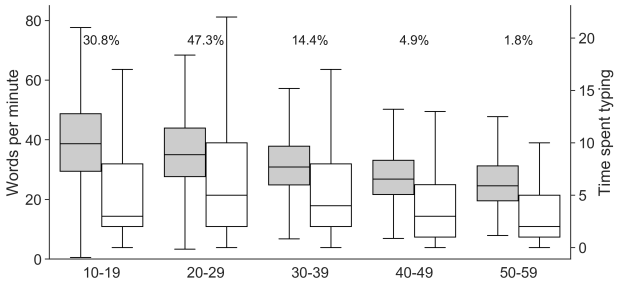

Figure 3: Performance (gray) and time spent typing on mobile devices (white) for different age groups with 95% confidence intervals and percentage of participants in each group.

Age: Participants' performance vary with age groups, as shown in Figure 3. Differences were significant for all groups (adj. $p < 0.001$). Participants of age between 10 and 19 typed the fastest ($M = 39.6$, $SD = 14.3$), participants of age below 10 – the slowest ($M = 24.3$, $SD = 13.2$). Interestingly, participants aged 10–19 were not the ones who reported the most time spent typing, as shown in Figure 3). Note that this analysis excludes 0.2% of participants older than 60.

English experience: Native users of English were the fastest typists ($M = 37.8$, $SD = 13.6$); those who never type in English the slowest ($M = 25.6$, $SD = 12.4$). Figure 4 shows how the the typing speed decreased with reported level of experience (adj. $p < 0.001$ for all comparisons except SOMETIMES versus RARELY, adj. $p = 0.0173$).

Typing training: Surprisingly, users who reported to have taken a touch typing course for typing on desktop keyboards were slightly slower ($M = 35.7$, $SD = 13.4$) than those who reported to not have taken such a course ($M = 36.4$, $SD = 13.1$). Note that this difference was significant, albeit with a small effect size (adj. $p < 0.001$, $d = 0.002$).

Finger usage: Participants who reported to use two fingers were significantly faster than those who used only one finger ($M = 37.7$, $SD = 13.2$ versus $M = 29.2$, $SD = 10.7$, $p < 0.001$, $d = 0.66$). A closer look at the reported typing posture shows that the use of different hands and fingers had a significant impact on performance. Over 82% of participants typed using two thumbs. Confirming the findings of prior work [3, 7], this was the fastest way to enter text ($M = 38.02$, $SD = 13.1$, $p < 0.001$ in comparison to all other groups). Those who typed with the index finger of the left hand were the slowest ($M = 25.0$, $SD = 11.4$, adj. $p < 0.001$ in comparison to all other groups but RIGHT, INDEX, adj. $p = 0.014$). Figure 4 shows the performance and frequency for all typing postures. Note, that this analysis excludes a small percentage of people (< 1%) who reported to use the middle finger(s).

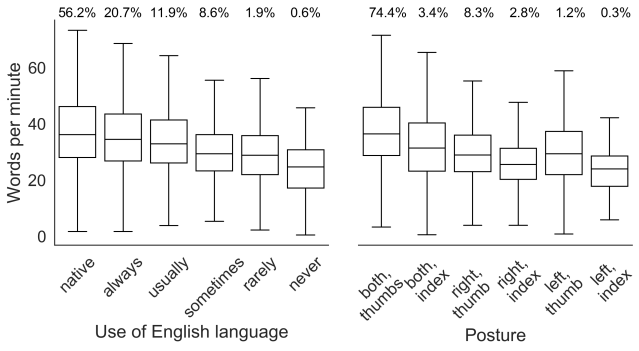

Figure 4: Typing speed versus use of the English language and posture in daily typing with 95% confidence intervals and percentage of participants in each group.

Discussion. The observed performance differences are large with over 12 WPM difference between native English speakers and those that never type in English. This holds important implications for text entry studies which are often performed with non-natives typing in English, but disregarding the language experience in the analysis. From the results above, we also conclude that a touch typing course on physical keyboards has no recognizable association with typing performance on mobile phones operated with only 1 or 2 fingers.

Intelligent text entry

Based on our ITE recognition method, we classified each input event into prediction, autocorrection, gesture, or none. We here analyze the actual use of ITE in practice and its correlation with typing behavior. For each ITE, we computed the percentage of words entered using the ITE per participant.

Usage and performance. We found 13.9% of participants did not use any ITE method. More than half of the participants used a mix of ITEs. Exact numbers are given in Figure 5. On average, across participants who used any of the ITEs, 8% of words were automatically corrected, 10% of words were selected from the prediction list, and 22% of words were entered using a gesture.

Impact of ITE on performance. The use of different ITE methods is associated with WPM and other performance measures. Figure 5 compares the WPM between each group. Participants that used autocorrection were faster ($M = 43.4$, $SD = 14.4$) than all other participants ($p < 0.001$ for all comparisons with other groups, using Holm-Bonferroni correction). Participants using prediction only or in combination with gestures were the slowest, with 10 WPM less than those using autocorrection. Pairwise-comparisons using Holm-Bonferroni correction showed significant differences between participants using no ITE and those using prediction (adj. $p < .001$, $d = .15$), a mix of prediction

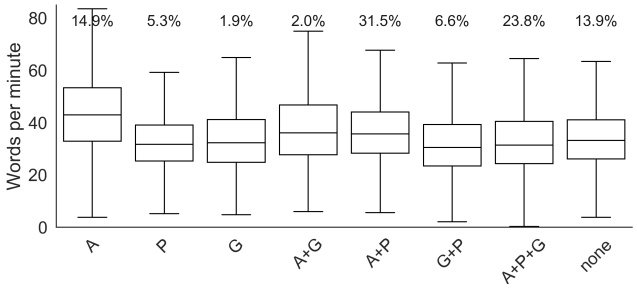

Figure 5: ITE method versus typing speed with 95% confidence intervals and percentage of participants in each group. P = Prediction; A = Autocorrection; G = Gesture.

and gestures (adj. $p < .001$, $d = .07$), or all ITEs (adj. $p = 0.04$, $d = .49$). These differences are less pronounced in the U.S. subsample where the difference between autocorrection and normal typing was not found to be significant, nor were the difference between prediction and normal typing. The detailed statistics can be found on the project page userinterfaces.aalto.fi/typing37k. Exact numbers of performance are given in Table 5.

A correlation analysis between the different ITE methods and performance metrics further confirms this observation. As shown in Table 6, autocorrection has a moderate positive correlation with WPM ($r = 0.237$). This is plotted in Figure 6c. Conversely, word prediction has a small negative correlation with performance ($r = -0.183$), as shown in Figure 6b. Our correlation analysis shows that the use of ITE affects KSPC. Using gestures and word prediction reduces the amount of keystrokes ($r = -0.251$ and $r = -0.232$, respectively). In contrast, autocorrection has a positive correlation with KSPC, indicating an increase in keystrokes. Similar effects are observed for the U.S. subsample, as shown in Table 6.

Table 5: Performance measures for each group of intelligent text entry methods and their combinations, overall and in the U.S. subsample. P=Prediction; A=Autocorrection; G=Gesture.

	Overall participants			U.S. subsample		
ITE	WPM (SD)	ER	KSPC	WPM (SD)	ER	KSPC
none	34.8 (12.6)	2.3	1.2	42.6 (14.7)	2.2	1.2
P	32.8 (12.1)	2.3	1.2	35.4 (15.1)	2.2	1.1
A	43.4 (14.4)	2.4	1.2	46.1 (14.4)	2.3	1.2
G	32.2 (13.4)	2.4	1.0	38.8 (12.5)	2.1	0.8
P+A	35.7 (12.6)	2.4	1.2	37.3 (12.8)	2.4	1.2
P+G	31.5 (13.6)	2.2	0.9	37.5 (19.3)	2.1	0.7
A+G	33.8 (12.1)	2.4	1.1	33.3 (10.9)	2.6	1.2
P+A+G	28.8 (11.3)	2.4	1.1	30.9 (13.4)	2.1	1.1

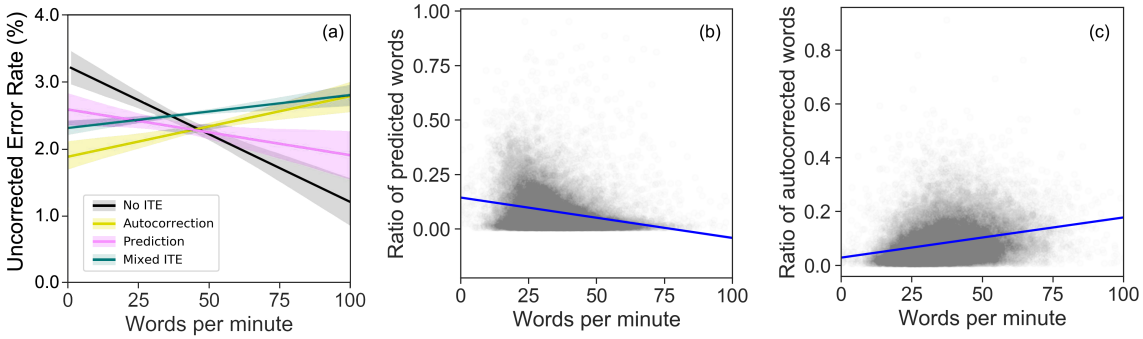

Figure 6: (a) Regression analysis for the relation between error rate and performance for different groups of ITE methods. The bands denote 95% confidence interval. Note that the Gesture group is not shown because of small sample size (<100). (b, c) Typing performance in relation to the use of intelligent text entry methods. There is (b) a negative correlation with percentage of words typed using prediction, and (c) a positive correlation with percentage of autocorrected words.

Impact of ITE on error. To analyze the effect of ITE methods on error and how it changed performance, we performed a regression analysis for each ITE condition as shown in Figure 6a. For participants using no ITE, we found a weak negative correlation between error rate and performance ($r = -0.16$). This means, without ITE, the faster typists tend to generate less error. This is in line with earlier findings on desktop typing [11]. In contrast, other groups do not show clear trends between the error rates and performance. Almost zero correlation were found for autocorrection ($r = 0.07$), prediction ($r = -0.05$) and mix of ITE ($r = 0.04$). Note that gesture input was not analyzed due to its small sample size.

Discussion. In comparison to what has been reported in a smaller-scale study that used a dedicated typing application and smaller convenience sample [7], users of our typing test used autocorrection and prediction for more words. This might be due to the keyboard being more familiar to them.

Prior work has noted that autocorrection can be detrimental to performance because of high cost of erroneous corrections [7]. It is interesting to see that nevertheless, participants using autocorrection have the highest performance in our dataset.

Table 6: Pearson correlation between performance and ITE measures, overall and in the U.S. subsample. Gray: not statistically significant ($p > 0.05$), bold: weak or moderate correlations. P=Prediction; A=Autocorrection; G=Gesture.

	Overall participants			U.S. subsample		
Measure	G	A	P	G	A	P
WPM	-0.003	**0.237**	**-0.183**	-0.017	**0.272**	**-0.152**
ER	-0.012	0.086	-0.037	0.006	0.051	-0.005
KSPC	**-0.251**	**0.181**	**-0.232**	**-0.219**	**0.171**	**-0.283**

As discussed at the beginning of the paper, prior work on the usefulness of word prediction has presented conflicting results. The performance benefit depends on many factors. For mobile typing, a recent study showed decreased performance rates for heavy use of word prediction [29]. Our correlation analyses reveal similar trends. However, the wide spread of data points in Figure 6 shows the need for more detailed analyses to better understand the usefulness of ITE in different contexts and for different users. While faster typists generally make fewer mistakes [11], we found no such relation in the case where ITE methods are used. This indicates that the use of autocorrection and prediction mitigates the higher error rate of novice users.

6 THE DATASET

We release a dataset containing typing events from over 37,000 participants. It includes all data reported on here, including demographics, key log data, stimuli and transcribed sentences, key press events and corresponding state of the input field. In addition, we captured each device's screen width and height, the device type and brand, the keyboard app as reported by the participants, as well as the device's orientation at every key press. The dataset and preprocessing code are available at https://userinterfaces.aalto.fi/typing37k.

7 DISCUSSION

In this work, we collected typing data from 37,370 volunteers using a browser-based transcription test. Previous work on gathering typing data outside the traditional lab experiment has relied on crowdsourcing [22] or custom mobile apps [7, 16, 31]. In contrast to previous studies, the dataset in this paper is on an unprecedented scale. However, this comes with limitations. Generalizability of the sample is an issue: our participants are likely exhibiting a self-selection bias due to the nature of the website, which is a typing test

website. Many participants are young females from the U.S. interested in typing. This is not representative of the general population and might bias the data towards representing a western, young, more technology-affine group of people. We compared our results to a subsample that better represents U.S. demographics and could not find any significant differences for the basic performance measures. Nevertheless, results might not be generalizable to other user groups. One example of likely sampling bias influence is in the low proportion of gesture keyboard users. Researchers interested in using this dataset for their research should first consider this sampling method and its limitation.

On the other hand, previous solutions for collecting mobile typing data typically relied on either opportunity-sampling from a university campus population or recruiting participants from microtask markets or app markets, which also introduces bias, though not necessarily in terms of the same factors. A fruitful avenue of future work would be to perform a factor analysis and identify the dominant user factors influencing typing performance and typing behavior. Such work could be used to correct sampling errors in text entry studies regardless of the participant recruitment source.

We observed a higher text entry rate in our sample for auto-correction and a lower entry rate for word prediction. The efficacy of word prediction is an open research problem as it depends on several factors. The primary one is the accuracy of word prediction and the unaided entry rate of the user, in other words, to which degree the user is rate-limited. We conjecture that the relatively high entry rates we observed overall in our sample make it challenging for word prediction to provide a substantial performance benefit for users. These results are in line with prior lab studies on mobile typing [29].

Note that our analysis is limited by the accuracy of the ITE recognition. Due to the security and privacy restrictions of mobile devices, we were often unable to log keycode information of each keypress. To detect and analyze the use of intelligent text entry methods, we had to rely on a heuristic recognition scheme based on changes in the input field. To evaluate our recognition we collected a ground-truth dataset from video recordings of 15 participants. We found that our technique was simple but effective, classifying over 90% of ITE events correctly with a low false-postive rate ($< 1\%$). However, there were a few cases where the changes in the input field were ambiguous and correct recognition was not always possible. Our evaluation study showed that for such edge cases, ITE use was not recognized resulting in false-negatives – the majority of misclassifications. We argue that our findings on the effect of ITE on performance should not be affected by this; if so effects should be even more pronounced. Future work could investigate the use of more advanced learning-based approaches to recognize ITE usage from changes in the input field.

We used a transcription task to assess typing performance which requires the participant to dedicate part of their attention to the transcribed sentence in addition to the entered text and the keyboard. It is also possible to use alternative methods, such as a composition task [38] or even object-based methods, such as instructing users to annotate an image [12]. Given the high traffic volume for the data tap underpinning this work, we see promising follow-up work in both changing the nature of the task and the parameters of individual tasks. For example, using this webpage, we could investigate the effect of different composition tasks and individual task parameters, such as the effect of difficulty of a sentence set [36] on transcription task performance. Such investigations are difficult to perform using traditional text entry experimental methods and we hope the data tap approach will be inspirational for other text entry researchers.

8 CONCLUSIONS

In this paper, we have reported observations from a transcription task mobile text entry study with 37,370 volunteers. The set-up allowed us to carry out detailed statistical analyses of distributions and correlates of typing performance, including demographics, device, and technique. Due to the size of the dataset, this paper has been able to reveal the distributions of key text entry metrics, such as entry rate, uncorrected error rate and keystrokes per character for both the entire sample and a stratified subsampled dataset designed to represent U.S. mobile text entry users. Also, we have classified the participants' typing into four different technique categories: autocorrect, word prediction, gesture keyboard, and plain typing. This allowed us to explore the effects of these techniques. Among other findings, the data indicates that autocorrect users tend to be faster while those that rely on prediction tend to be slower. The collected dataset is very rich. The presented analysis confirms prior findings on smaller more controlled studies and gives us new insights into the complex typing behavior of people and the large variations between them. However, more research is needed to disentangle confounds, and to investigate other factors and their interactions. To this end, we are releasing the code and the dataset to assist further efforts in modeling, machine learning and improvements of text entry methods.

9 ACKNOWLEDGEMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 637991), the ERC Grant OPTINT (StG-2016-717054) and EPSRC (grant EP/R004471/1). The data collection was supported by Typing Master, Inc.

REFERENCES

[1] Denis Anson, Penni Moist, Mary Przywara, Heather Wells, Heather Saylor, and Hantz Maxime. 2006. The Effects of Word Completion and Word Prediction on Typing Rates Using On-Screen Keyboards. *Assistive Technology* 18, 2 (sep 2006), 146–154. https://doi.org/10.1080/10400435.2006.10131913

[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)*. ACM, New York, NY, USA, 415–420. https://doi.org/10.1145/2992154.2996791

[3] Shiri Azenkot and Shumin Zhai. 2012. Touch behavior with different postures on soft smartphone keyboards. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services - MobileHCI '12*. https://doi.org/10.1145/2371574.2371612

[4] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources?. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. 356–364.

[5] Nikola Banovic, Varun Rao, Abinaya Saravanan, Anind K. Dey, and Jennifer Mankoff. 2017. Quantifying Aversion to Costly Typing Errors in Expert Mobile Text Entry. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4229–4241. https://doi.org/10.1145/3025453.3025695

[6] Xiaojun Bi, Yang Li, and Shumin Zhai. 2013. FFitts Law: Modeling Finger Touch with Fitts' Law. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1363–1372. https://doi.org/10.1145/2470654.2466180

[7] Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, New York, New York, USA, 1–14. https://doi.org/10.1145/3173574.3173829

[8] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[9] Pew Research Center. 2015. U.S. Smartphone Use in 2015. http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/

[10] countrymeters.info. 2019. Live United States of America (USA) population. https://countrymeters.info/en/United_States_of_America_(USA) accessed January 30, 2019.

[11] Vivek Dhakal, Anna Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. Observations on Typing from 136 Million Keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, New York, USA. https://doi.org/10.1145/3173574.3174220

[12] Mark D. Dunlop, Emma Nicol, Andreas Komninos, Prima Dona, and Naveen Durga. 2016. Measuring inviscid text entry using image description tasks. In *Proceedings of the 2016 CHI Workshop on Inviscid Text Entry and Beyond*.

[13] M Goel, A Jansen, T Mandel, S Patel, and J Wobbrock. 2013. ContextType: using hand posture information to improve mobile touch screen text entry. In *Proceedings of the 31st ACM international conference on Human factors in computing systems CHI '10*. https://doi.org/10.1145/2470654.2481386

[14] Joshua Goodman, Gina Venolia, Keith Steury, and Chauncey Parker. 2002. Language Modeling for Soft Keyboards. In *Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI '02)*. ACM, New York, NY, USA, 194–195. https://doi.org/10.1145/502716.502753

[15] Asela Gunawardana, Tim Paek, and Christopher Meek. 2010. Usability guided key-target resizing for soft keyboards. In *Proceedings of the 15th international conference on Intelligent user interfaces - IUI '10*. ACM Press, New York, New York, USA, 111. https://doi.org/10.1145/1719970.1719986

[16] Niels Henze, Enrico Rukzio, and Susanne Boll. 2012. Observational and experimental investigation of typing behaviour using virtual keyboards for mobile devices. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. ACM Press, New York, New York, USA, 2659. https://doi.org/10.1145/2207676.2208658

[17] indexmundi.com. 2019. United States Age structure - Demographics. https://www.indexmundi.com/united_states/age_structure.html accessed January 30, 2019.

[18] Heidi Horstmann Koester and Simon Levine. 1996. Effect of a word prediction feature on user performance. *Augmentative and Alternative Communication* 12, 3 (jan 1996), 155–168. https://doi.org/10.1080/07434619612331277608

[19] Andreas Komninos, Mark Dunlop, Kyriakos Katsaris, and John Garofalakis. 2018. A glimpse of mobile text entry errors and corrective behaviour in the wild. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct - MobileHCI '18*. ACM Press, New York, New York, USA, 221–228. https://doi.org/10.1145/3236112.3236143

[20] Per Ola Kristensson. 2007. *Discrete and continuous shape writing for text entry and control*. Ph.D. Dissertation. Institutionen för datavetenskap.

[21] Per Ola Kristensson. 2009. Five challenges for intelligent text entry methods. *AI Magazine* 30, 4 (2009), 85.

[22] 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*. ACM, 29–32.

[23] Per-Ola Kristensson and Shumin Zhai. 2004. SHARK2: a large vocabulary shorthand writing system for pen-based computers. In *Proceedings of the 17th annual ACM symposium on User interface software and technology - UIST '04*. ACM Press, New York, New York, USA, 43. https://doi.org/10.1145/1029632.1029640

[24] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. 707–710.

[25] I Scott MacKenzie and R William Soukoreff. 2002. A character-level error analysis technique for evaluating text entry methods. In *Proceedings of the second Nordic conference on Human-computer interaction*. ACM, 243–246.

[26] I. Scott MacKenzie and Shawn X. Zhang. 1999. The design and evaluation of a high-performance soft keyboard. In *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*. ACM Press, New York, New York, USA, 25–31. https://doi.org/10.1145/302979.302983

[27] Josip Musić and Roderick Murray-Smith. 2016. Nomadic Input on Mobile Devices: The Influence of Touch Input Technique and Walking Speed on Performance and Offset Modeling. *HumanâĂŞComputer Interaction* (2016). https://doi.org/10.1080/07370024.2015.1071195

[28] Antti Oulasvirta, Anna Reichel, Wenbin Li, Yan Zhang, Myroslav Bachynskyi, Keith Vertanen, and Per Ola Kristensson. 2013. Improving two-thumb text entry on touchscreen devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. ACM Press, New York, New York, USA, 2765. https://doi.org/10.1145/2470654.2481383

[29] Philip Quinn and Shumin Zhai. 2016. A Cost-Benefit Study of Text Entry Suggestion Interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, New York, New York, USA, 83–88. https://doi.org/10.1145/2858036.2858305

[30] Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1364–1378. https://doi.org/10.1145/2675133.2675246

[31] 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

[32] Sayan Sarcar, Jussi PP Jokinen, Antti Oulasvirta, Zhenxin Wang, Chaklam Silpasuwanchai, and Xiangshi Ren. 2018. Ability-Based Optimization of Touchscreen Interactions. *IEEE Pervasive Computing* 17, 1 (2018), 15–26.

[33] I Scott MacKenzie and R William Soukoreff. 2002. A Model of Two-Thumb Text Entry. *Graphics Interface* (2002), 117—-124. http://www.graphicsinterface.org/wp-content/uploads/gi2002-14.pdf

[34] Keith Trnka, John McCaw, Debra Yarrington, Kathleen F. McCoy, and Christopher Pennington. 2009. User Interaction with Word Prediction. *ACM Transactions on Accessible Computing* 1, 3 (feb 2009), 1–34. https://doi.org/10.1145/1497302.1497307

[35] Paul D. Varcholik, Joseph J. LaViola, and Charles E. Hughes. 2012. Establishing a baseline for text entry for a multi-touch virtual keyboard. *International Journal of Human-Computer Studies* 70, 10 (oct 2012), 657–672. https://doi.org/10.1016/J.IJHCS.2012.05.007

[36] K. Vertanen, D. Gaines, C. Fletcher, A.M. Stanage, R. Watling, and P.O. 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 - CHI '19*. ACM Press, New York, New York, USA, to appear.

[37] 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*. ACM, 295–298.

[38] Keith Vertanen and Per Ola Kristensson. 2014. Complementing text entry evaluations with a composition task. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 2 (2014), 8.

[39] 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 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 659–668.

[40] Daryl Weir, Simon Rogers, Roderick Murray-Smith, and Markus Löchtefeld. 2012. A User-specific Machine Learning Approach for Improving Touch Accuracy on Mobile Devices. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST '12)*. ACM, New York, NY, USA, 465–476. https://doi.org/10.1145/2380116.2380175

[41] Jacob O. Wobbrock. 2007. Measures of Text Entry Performance. In *Text Entry Systems*, Scott I. MacKenzie and Kumiko Tanaka-Ishii (Eds.). Morgan Kaufmann, 47–74. https://doi.org/10.1016/B978-012373591-1/50003-6

[42] Jacob O. Wobbrock and Brad A. Myers. 2006. Analyzing the Input Stream for Character- Level Errors in Unconstrained Text Entry Evaluations. *ACM Trans. Comput.-Hum. Interact.* 13, 4 (Dec. 2006), 458–489. https://doi.org/10.1145/1188816.1188819

[43] Ying Yin, Tom Y. Ouyang, Kurt Partridge, and Shumin Zhai. 2013. Making Touchscreen Keyboards Adaptive to Keys, Hand Postures, and Individuals - A Hierarchical Spatial Backoff Model Approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/2470654.2481384

[44] Shumin Zhai and Per-Ola Kristensson. 2003. Shorthand writing on stylus keyboard. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 97–104.