Mobile Typing with Intelligent Text Entry: A Large-Scale Dataset and Results

Katri Leino, Markku Laine, Mikko Kurimo, Antti Oulasvirta

Department of Information and Communications Engineering, Aalto University, Espoo, Finland.

Contributing authors: katri.k.leino@aalto.fi; markku.laine@gmail.com; mikko.kurimo@aalto.fi; antti.oulasvirta@aalto.fi;

Abstract

A large proportion of text is produced using mobile devices. However, very little research looks at the special characteristics of how this happens and, importantly, how it is affected by the design of the language model (LM). The operating systems of modern devices offer a number of LM-based intelligent text entry methods (ITEs) such as Autocorrection (AC) and Suggestion Bar (SB) to aid typing. It is not known how the keyboard and performance of the ITEs influences the typing strategies in the wild. LMs are operating system and language-specific, therefore, in this paper, we release and analyse a large-scale dataset of mobile typing in two languages: English (46 755 participants) and Finnish (8661 participants). Typing data was collected with the participants' own iPhone and Android devices resulting a diverse data on the ITE method performance. By analysing the typing speed and the information on which letters and operations happen in each keystroke, we found that iPhone and Android devices encourage the use of two different typing styles. IPhone users used utilise mainly AC and are able to achieve the highest typing speed among the participants when the balance between AC accuracy and threshold to correct user error are adequately balanced. Android users prefer SB to avoid and correct typing errors instead of utilising AC. Especially Finnish participants, who had low ITE accuracy, used SB often to correct typing errors. To develop and evaluate LMs for typing applications, it is essential to study which factors affect the user. The typing dataset that we have prepared and opened for the public, allows for analysing the factors thus aiding the development of the most useful LMs.

Keywords: intelligent text entry, mobile typing, autocorrection, suggestion bar, language models

1 Introduction

Intelligent Text Entry (ITE) methods are the central solution to typing accuracy issues and to improve the typing experience on mobile devices. The lack of tactile feedback and small key size makes typing slower and less accurate compared to physical keyboards. The two most popular ITE methods are language model-based Autocorrection (AC) and Suggestion Bar (SB) which are included in commercial smart devices at least for the most common languages. AC aims to correct the user's typing errors automatically after the word is completed. In SB, the user completes or corrects the word by selecting an option from the Suggestion Bar which is located above the keyboard. ITE methods are built on language models (LMs) such as N-gram models [1] or complex deep neural network models trained with a large amount of text data [2–4]. LMs can be used to calculate the most likely words in a certain context making them ideal for error correction and prediction tasks. In mobile devices, the LM is provided by the operating system and its architecture and performance are not public information. The only way to study how LM characteristics influence the typing strategies in the wild is to observe the ITE performance during the typing.

The effect of AC and SB on typing has been studied before either with computational user modelling [5] or in a controlled laboratory environment with a small amount of participants and devices. To control the influencing factors, custom keyboards or simulated ITE methods are often used. The keyboard design [6] and the accuracy of the ITE methods [7] affect how people type. However, in literature, the ITE methods are still typically studied without paying attention to operating systems, language and the LMs which have a direct effect on the typing experience. The development of the LMs is focused on the most common languages such as English, thus the performance of ITE methods on less-resourced languages is often lacking. The quality of the LMs is an important aspect to study as poor accuracy of the ITE methods may frustrate the user leading the user to stop using them [8]. Therefore, to develop and evaluate LMs for typing applications, it is essential to study which factors affect the user. In this paper, we focus on the ITE performance on the most common commercial devices which are used daily by the user. To study how ITE methods affect typing in real life, we extended the previously gathered large real-world English typing dataset [9] with additional participants, keystroke-level labels and collected similar typing data in Finnish. In the typing dataset, 46 755 English and 8661 Finnish participants each typed 15 sentences randomly selected from a set of sentences (EN:1525 and FI:550 sentences).

The main contribution of this paper is adding robust keystroke-level labels for the AC and SB operations to the typing data. In the keystroke-level dataset, each keystroke is labelled by added and deleted letters, and by additional operations such as ITE methods or system inputs. The automatic labelling supports various devices and can account for the additional keystrokes in the data caused by the device instead of the user. The structure of the data is visualised in Table 1. The dataset contains four data tables: (1) participant information, (2) typed sentences, (3) autocorrected and selected words, and (4) keystrokes. The new additions to the previously published data

Table 1: The typing dataset consists of four data tables representing different levels of data: participants, test sentences, ITE words and the log of the user's keystrokes. The data, added metrics and labels contained in each table is listed under the name of the table. The Finnish part is completely new and the new content of the English part compared to the previous version [9] is in *cursive blue.*

Structure of the typing dataset

Participants	Test sentences	*ITE Words*	Keystroke log
Questionnaire Device	Sentence Use input	*Typed word* *AC word* *Suggested word*	Timestamp **Input from device**: Key, Key code **User input:** Current text, *Added and deleted letters* *Space, BS* ***Other labels:*** *AC, SB* *Auto-input*
Evaluation metrics: WPM, KSPC, BSPC, UER ***ITE Evaluation metrics:*** *ACPW, PPW* *Error rate, Correction rate*			

are stated in cursive. The raw and processed dataset[1] and codes for data collection, processing, labelling and analysis with documentation are publicly available[2]. The codes and documentation enable replicating the results and labelling of new datasets.

The dataset allows us to observe how the use of ITE methods changes along with the operating system and the typed language. It is also possible to study, for example, the effect the utilisation rate and error rate of the ITE methods have on the typing speed. Among the participants, there is a clear difference in typing strategies between iPhone and Android users. While almost all iPhone users utilise AC, Android users prefer filling or correcting words by selecting them from the Suggestion Bar. The difference between iPhone and Android AC is that the utilisation and error rate of AC is higher on iPhone devices. The English iPhone users can achieve the fastest typing speed among the participants when they rely on AC actively even though AC makes more incorrect corrections on average. It seems that the iPhone have in English found a good balance between the AC error rate and the threshold for correcting potential errors. However, among Finnish participants, AC is not popular most likely because the error rate of AC is too high and AC users have lower average typing speed compared to those who type without ITE. On Android devices, Finnish participants use SB more often to correct typing errors instead of AC.

In this paper, we make the following contributions: (1) A large keystroke-level labelled typing dataset in two languages: English and Finnish, and (2) analysing the effects AC and SB have on typing. We also publicly share all the resources: datasets, scripts and codes.

[1] https://doi.org/10.5281/zenodo.12528163
[2] https://github.com/aalto-speech/ite-typing-dataset

2 Related Work

As mobile devices have become more common, more often people use them to communicate with others by typing. Touch screens are the most common design choice for mobile devices due to the screen's adaptability to new user interface designs. However, typing is more challenging on the touch screen due to the small screen, and lack of tactile. It is more difficult for the user to visualise the exact location finger touches the screen [10]. Intelligent text entry methods have been developed to aid users in typing. In this section, we introduce the related work of mobile typing and ITE methods. Finally, we discuss the type of datasets that have been used in the previous studies.

2.1 Typing on mobile device

The keyboard on mobile touchscreen devices is smaller than the traditional physical keyboard and lacks tactile keys. Where on the physical keyboard, users typically type with multiple fingers and can type without looking at the keyboard [11], the mobile keyboard is more often used with one or two fingers [9] and typing requires constant attention shifts between the keyboard and text entry field [6]. Thus, mobile typing is slower and the typing accuracy is lower [12]. In addition to the increase in typing errors, noticing the errors is delayed because the touch keyboard user shifts attention constantly between the keyboard and the text entry field. Because the cost of error is higher for mobile keyboards, the users type slower to reduce the risk of error [7]. Even increasing the number of fingers does not improve the cost of error as while typing with two thumbs is faster, noticing typing errors is delayed compared to typing with a single finger [13] To counter the challenges, touch keyboards utilize language model-based ITE methods to correct errors or fill words during typing. The typing speed is a trade-off between speed and accuracy. Therefore, users adjust their typing strategy to find a balance between speed, accuracy and utilization of ITE methods which suit their capabilities [14].

2.2 Intelligent Text Entry Methods

The purpose of Intelligent Text Entry methods is to help users correct typing errors, fill in the words or provide alternative ways to type. These methods are built on Language Models which model the structure of the language and can predict the most common word sequences [4]. Language models used in commercial ITE methods are large and complex deep neural networks which have been trained with large text corpora [2, 3]. Most smartphones have either iOS (iPhone) or Android operating systems and both have developed their own Language models. How the models on iPhone and Android devices are built is not public information, thus studying the performance is only possible by observing the typing. The performance of the ITE on different operating systems in user experiments is not typically taken into account since the experiments typically use custom keyboards or a small number of different devices. With our dataset, it is possible to study how the typing strategies change based on the operating system and compare how ITE methods perform differently on different devices.

Various ITE methods exist on mobile devices. The most common methods are AC and SB which are the focus of this paper. Other relatively popular ITE methods are Gesture (also known as Swipe) [15], where the user writes a word by sliding the finger from letter to letter on the keyboard, and post-correction where the user presses the already typed word and is presented a list of correction to select from. Post-correction is more often utilised during post-editing texts [16, 17].

AC is a common ITE method on touch keyboards. In previous studies, AC has been found to improve the typing speed [9, 18]. AC users can increase their typing speed because typing is a trade-off between speed and accuracy. When they type while relying on AC, they can type without worrying about typing errors, thus improving their typing speed [5]. Therefore, AC users balance between speed and accuracy to optimize their typing. One of the key factors which affect the user's decision to utilize AC is the accuracy and the cost of error of the AC. Correcting AC errors takes more effort and time because errors are more difficult to notice and, therefore, the user decreases their typing speed to make fewer typing errors if AC has too high error rate [18, 19]. when correcting AC errors takes more effort and time. AC errors frustrate users [20] which can cause the users to stop using it [8]. However, even if the accuracy is perfect users do not want to give full control of error correction to the AC [5, 21]. In this paper, we explore how AC is utilised on a variety of devices and if we can replicate the results of the effects of the error and utilisation rate in the lab setting.

SB shows the user a list of words above the keyboard which the user may select to complete the word they are typing. In addition to word suggestions, whole phrases can also be suggested to the user [22, 23]. The SB updates the suggestions while typing, thus, regular use requires the user to switch attention from typing to the suggestions which slows the typing on average [6, 9, 19, 24] even though typing uses fewer keystrokes [19, 25]. In addition to attention shift, the selection process itself is also slow. To select a suggestion the user has to first read the provided suggestions, then evaluate the options and decide whether to select one or not. Despite its flaws, many users still chose to use SB. Quinn et. al. [24] found that the more suggestions were provided for the user, the more often the user used SB even though they did not benefit speed-wise. However, the non-native speakers [23] and the users with slower typing speed [22] benefit more from suggestions compared to other typists. Since SB has been previously studied in the lab setting, our dataset provides insight into how often and for which kind of words users chose to use SB on their own devices which helps us to understand why users choose to use SB even if it slows the typing down.

2.3 Typing Datasets

Mobile typing and ITE methods have been studied previously with user experiments, modelling and gathering typing data. Most of the typing research has been conducted with a user experiment in a controlled laboratory. The used device is often provided by the organizer and may use a custom keyboard. The laboratory environment enables tracking typing and attention shifts by tracking eye [13] and finger movement [11, 13] to closely monitor how the participant interacts with the device. In a controlled environment, it is easier to observe certain aspects of the typing by limiting external influences. While the laboratory environment provides detailed, good-quality data,

the sample size of the experiments is typically small [26] and participants have often similar backgrounds. In addition, the data collected by the user experiments is rarely publicly available and the typing does not necessarily reflect how participants would type normally if an unfamiliar device or keyboard is used in the experiment.

Typing has also been studied by modelling typing with computational models [5, 6]. Modelling provides a way to study user behaviour and adaptability in a controlled environment and can also be used to provide typing data. The models can simulate the physical and cognitive aspects of the user, however, the ITE methods on the models are simulated with heuristics and do not represent how methods would perform on a real device.

Online studies can gather large amounts of data from participants of a variety of backgrounds and devices. When the user experiment is conducted online, it is required either to set up a web application on a server or use crowd-sourcing platforms. The study may gather a large and diverse set of participants, however, it comes with a cost that the user's typing behaviour is only observed based on a questionnaire, what is shown on the input field and possibly the keystroke information (which is not accurate in all cases) given by the device. The challenge with online studies is the large drop-out rate, noisy data from the keyboard, uncertainty with the information provided by participants and the internal state of the participants. To control the keyboard and to get more robust detailed data from the user is possible with custom keyboard [15, 27]. However, the typing strategy changes with the keyboard. When a custom keyboard is used, ITE methods are also custom made [15, 19] and thus they do not perform exactly as the user is used to on their own device. Because the users adapt to the keyboards to optimize the typing experience [6], the typing changes when the keyboard or ITE method changes. The people who have used their own devices for a long time and have developed typing strategies which fit the best to their own preferences, the keyboard layout and the performance of ITE. Therefore, it is important to also analyse how users type and utilize ITE methods on their everyday devices. Because there is a lot of variation in devices such as size and operating system, a large and versatile dataset is needed for robust estimates.

To analyse typing or ITE performance requires timestamped keystroke-level data. Typing datasets is rarely publicly available which is problematic as gathering the data requires a lot of resources and without the data, the studies are difficult to replicate. Large, publicly available datasets containing ITE methods are

- Mobile typing dataset by 170 participants [19]. Participants typed with a custom keyboard which included Autocorrection and Suggestion Bar methods. Language: English.
- Gesture typing dataset by 1338 participants [15]. Data was gathered with a custom keyboard and has only gesture typing. Language: English.
- Mobile typing dataset by 37 000 participants [9]. Typing data consists of multiple ITE methods (Autocorrection, Gesture, Suggestion Bar) and devices but without robust keystroke-level labelling. Language: English.

This study is a follow-up to previous large-scale online studies of desktop-typing [28] and mobile typing [9]. The before-mentioned studies collected typing datasets

with online study and collected a large amount of typing data on various devices and users with different backgrounds. We expanded the mobile typing dataset [9] with English participants and collected typing data also in Finnish. Because the typing dataset [9] was collected with various devices and custom keyboards were not used, the dataset is noisy and tracking the use of ITE methods is difficult. The dataset was previously lacking robust ITE labels, which are necessary for detailed analysis of ITE methods. The main contribution of our paper is adding detailed labels for keystrokes and distinguishing keystrokes caused by the device (auto-inputs) and those made by the user. This level of labelling is not available on other publicly available datasets. Because the typing research mainly focuses on English, to our knowledge, the effects of the language on ITE methods have not been reported before.

3 Data Collection

The typing dataset was collected with a web-based transcription task on publicly available websites for English[3] and Finnish[4]. The websites support the popular mobile operating systems and browsers. The typing test on the websites holds a task to transcribe the given sentences and answer a questionnaire related to their typing. In this section, the collection procedure of the typing dataset is described focusing on the Finnish data collection. The collection implementation described here is an improved version of the implementation presented in the article from Palin et. al. [9] which also describes the collection of English typing dataset.

3.1 Participants

Participants volunteered in the study by taking an online Finnish mobile typing test between September 2019 and June 2020. The study was promoted in a Yle news article on mobile typing[5], on social media, and through mailing lists by researchers. The new version of the English dataset was collected between September 2018 and June 2020.

The studies collected data on 399 833 voluntary English participants and 22 082 Finnish participants. After the preprocessing (see Section 4.1), the number of subjects was reduced to 46 755 English and 8661 Finnish participants. For the analysis we only selected native speakers, thus, the resulting number of participants in the analysis subset was 24 750 English and 8481 Finnish participants. The demographics of the participants are presented in Section 5.1.

3.2 Task

The typing test contains a transcription task, where the participant is asked to type 15 sentences that were randomly sampled from a set of 550 sentences. Sentences are shown to the participants one at a time, and they are asked to first read the sentence carefully before typing it as accurately as possible. The sentence is always visible during the task. Participants' average typing speed and uncorrected error rate are

[3] https://typingtest.aalto.fi
[4] https://kirjoitustesti.aalto.fi
[5] Yle news article on mobile typing, https://yle.fi/aihe/artikkeli/2019/12/03/kuinka-sujuvasti-kirjoitat-kannykalla-kannykan-napyttelynopeudella-on-yhteys

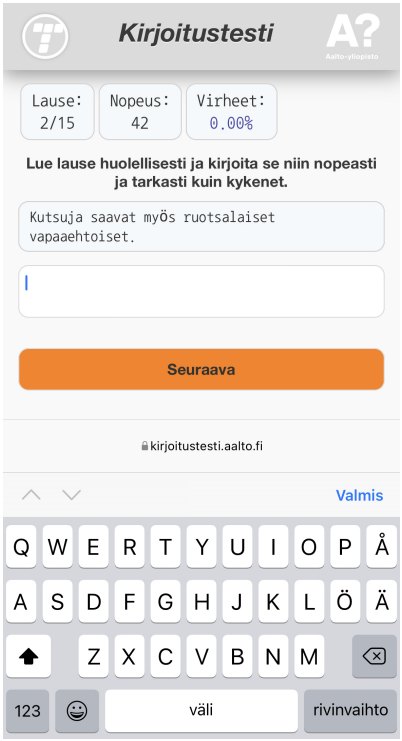


Fig. 1: The Finnish typing test website.

measured and reported to the user after each sentence. While showing typing speed to the participants might change their typical typing strategy, its function is to motivate the participant to complete the test properly. The typing test interface is shown in Figure 1. Participants used their own devices to complete the task.

After typing all the sentences, the participant fills out a questionnaire about basic information such as age and native language, and about the typing habits, i.e., used fingers and ITE tools in typing. Afterwards, the participant is given feedback by reporting the average WPM and uncorrected error rate in comparison to other participants in general and within their age group. Participants were also shown their fastest and slowest sentence and the sentence which had the highest error rate.

3.3 Questionnaire

In the questionnaire, the participant fills in age, gender and native language. They are also asked if they have taken any typing courses, which keyboard they are using, whether they are using any intelligent text entry methods and which fingers they use to type. In addition, we inquired how much time they type daily and how often they type the test's language.

3.4 Material

The Finnish phrase set contains 550 sentences sampled from two corpora: 266 phrases from Yle News Archive Easy-to-read Finnish 2011-2018 [29] and 284 phrases from Suomi24 online discussions [30]. Phrases were selected to represent the language often used daily. News phrases are formal Finnish used in official documents and written communication. Suomi24 is an online discussion platform, and the Suomi24 dataset contains phrases from real forum discussions. Phrases are written in colloquial Finnish which, compared to formal Finnish, has different vocabulary and sometimes different spelling as the words are often shortened or combined. The challenge of mixing formal and conversational language in the study was that participants may more easily mistype the conversational language if they are expecting formal Finnish vocabulary. Language models used in ITE often struggle with colloquial Finnish, therefore it was important to be included in this study.

Sentences were selected with restricted random sampling. The random sampling was restricted by the length of the sentence, the length of the words and the frequency of the word in the corpus. The restrictions aimed to select simple and short sentences which are easy to read and remember so that participants can focus on typing. We also wanted both phrase sets to be similar. The sentences were scored to select sentences with a certain length and difficulty. The difficulty depends on the length and rarity of the words in the sentence. Phrases containing short and common words are the fastest to type [31]. The sentences were scored by multiplying the length of the word $len(word)$ with a logarithm of the word's frequency $freq(word)$ and summing scores of words in the sentence.

$$\sum_{i=0}^{N} \text{len}(word)_i \times \log \text{freq}(word)_i \,, \tag{1}$$

where N is the total number of words in the sentence. After the sampling, inappropriate sentences (e.g. containing swearwords) and sentences containing difficult names were removed.

In the Finnish typing test, we aimed for sentences which are comparable to the sentences in the English typing set. The phrase set of the English typing test contained 1525 sentences which were random sampled from the memorable set of Enron mobile email corpus [32] and English Gigaword Newswire dataset [33]. The same phrase set has been used in previous studies [9, 28]. The Finnish phrase set has fewer sentences than the English set due to the expectation that the Finnish study would gather fewer participants.

3.5 Implementation

We implemented *Kirjoitustesti* as a web application consisting of a frontend (user interface) and a backend (server). The frontend is built with standard web technologies (HTML, CSS, JavaScript) and popular third-party frameworks and libraries, such

as Bootstrap[6] and jQuery[7]. The backend uses the Play Framework[8] for Scala and MySQL[9] database to compute analyzed metrics to be stored along with collected typing test data.

The device, the size and orientation of the device, the operation system, and the browser participants used were saved to the server. During the test, each keystroke was saved with a timestamp (keyup and keydown), the current text in the input field, and a keycode.

4 Preprocessing and Labelling Data

The standard procedure before dataset can be used, is to clean the dataset by removing incomplete and unnecessary information. In this section, we explain how the typing data was preprocessed and labelled with performance measures and keystroke-level labels.

4.1 Preprocessing

The raw dataset is processed to remove incorrect and incomplete samples and to create a suitable and consistent dataset to study the effects of the chosen ITE methods. We selected only mobile users between the ages of 10 and 70 who completed at least 15 sentences and answered the questionnaire on either Android or iPhone devices. Participants had the option to complete more than 15 sentences, but only the first 15 sentences were included in the dataset so that participants are comparable to each other. 16.8% of Finnish participants and 24.8% of English participants completed more than 15 sentences. The average WPM of the user is required to be over 0 but under 200. It was also required that the edit distance between the typed sentence and the original sentence be less than 25 characters. We limited our study only to native speakers because the Finnish dataset has only 2.2% non-native participants while 47.3% of the English participants were non-native speakers. Natives type differently compared to those who do not type in their native language [34].

To study how AC and SB affect typing, we only select users who have reported either using AC, SB, or both of them or no ITE at all. Gesture (Swipe) users were left outside of this study because Gestures are difficult to distinguish from SB suggestions during the automatic labelling. We also filtered out users whose input was not logged correctly in the keystroke level by setting that keystroke per character (KSPC) should be more than 0.5 and AC and SB cannot be used more than for 70% of the words. If AC or SB is used more often, most likely there is an error in the keystroke level data and, therefore, in the labelling.

Preprocessing reduced the number of English participants from 399 833 to 24 750. With non-natives, the number of participants is 46 755. The Finnish dataset had initially 22 082 participants and after preprocessing 8481.

[6] Bootstrap, https://getbootstrap.com/
[7] jQuery, https://jquery.com/
[8] Play Framework, https://www.playframework.com/
[9] MySQL, https://www.mysql.com/

4.2 Performance measures

We compute additional labels to evaluate the typing performance for each typed sentence and participant [35].

Word Per Minute (WPM): measures how many words are typed in a minute. It is the most common measure of typing speed. A formula for computing WPM is the following

$$WPM = \frac{|T| - 1}{S} \times 60 \times \frac{1}{5}, \tag{2}$$

where $|T|$ is the length of the character string and S is the time in seconds between the first and the last keystroke. Word length is assumed to be 5 for both languages so that the results are comparable. [35]

Keystrokes per character (KSPC): The number of keystrokes divided by the number of characters in the sentence. Measures how many keystrokes are required for a single character.

Backspaces per character (BSPC): The number of backspace presses in the test section. For the overall average, the number of backspaces is normalized by the number of characters in the sentence.

4.3 Labelling data

Keystroke-level data contains information on each keystroke logged during the test. During each keystroke, key codes from the device and the current input are recorded into the dataset. Because key codes are often unidentified due to differences in devices, the labelling is based on the text input field. We label each keystroke with information on which letters were added or removed. Table 1 lists the added labels. Next, we explain more closely how the data was labelled.

In general, AC and SB can be recognized by the following descriptions [9]:

Autocorrection (AC): AC automatically corrects word when a user presses a space. A keystroke is recognized as AC if more than one character is changed and the last character is a space key or punctuation. Automatic capital letter corrections are also counted as AC. Auto-fill is ITE method where user completes a unfinished word by pressing a space during the typing. Auto-fill suggestion is typically presented above the word currently being typed and needs to be rejected if not wanted. AC and auto-fill are difficult to distinguish based on text input only. Therefore, we do not attempt to separate the two in this study.

Suggestion Bar (SB): The user selects a word from the list of words which are presented above the keyboard. The selected word then replaces the word that was being typed. A keystroke is recognized as a SB selection if more than one character is changed and the last letter is not space or punctuation.

The descriptions above are the main guidelines. The challenge with the keystroke data is that it is device-dependent. In practice, keystroke data contains additional keystroke logs from inputs caused by the touch keyboard. On some devices, the AC and SB inputs do not abide by the description, which causes inaccurate labelling. Since participants use their own devices, it is not possible to account for all varieties

in the data. However, we were able to take into account the most common variants which are explained more thoroughly in the code documentation.

The selection of a SB is often accompanied by some sort of auto-input as the word is replaced by a word selection. We call the keystrokes caused by the device auto-inputs because they are not caused directly by the user pressing a key but by the touch keyboard itself. Auto-inputs can usually be recognized by that they are faster (10-25ms) than human keystrokes. One of the common variants of a SB is that when the user selects a word from the list, the old word disappears completely before the selected word appears. Due to these exceptions SB recognition is not trivial, and can be often confused with Gesture (Swipe) inputs as they in some cases have identical representations in the keystroke data.

Another challenge is that other automatic corrections can be mislabelled as AC or Suggestion. For example, the code can identify automatic correction when two spaces transform into a dot on iPhone devices, and when additional space is removed after AC if punctuation is typed.

The accuracy of ITE methods is recognized by comparing the SB selection or autocorrected word to the original word. We also track ITE words in case the user does not correct the error immediately. Users may correct words with backspaces or reverse the autocorrected or SB selected word to the previous state.

To evaluate the performance of the AC and SB, we added the following measures to participant and sentence-level data based on the labels on the keystroke-level data.

Autocorrection per word (ACPW): The number of times AC occurred in the sentence normalized by the number of words in the sentence.
Suggestion Bar selection per word (SBPW): The number of times the word from the SB is selected while typing the sentence is normalized by the number of words in the sentence.
ITE accuracy: Accuracy of AC or SB; how often the autocorrected or selected word was correct.
ITE correction rate: How often the user corrected the word after either AC or SB was used.

5 Typing dataset statistics

In this section, we summarize the demographics of the participants subgroup whose typing is analysed in the next section and compare the differences between the self-reported and the observed ITE methods. Lastly, we present the ITE related statistics.

5.1 Participants

The participants consist of only native speakers of each language. As is listed in Table 2, the subset used for the analysis contains 24 750 English participants and 8481 Finnish participants. The average age of Finnish participants is 28.8 (std:11.7) years which is 2.9 years older than English participants with an average age of 25.9 years (std:10.6). While most of the participants were between the ages of 15 and 35 years, both datasets have a reasonable number of older participants: the English set has 2651

and the Finnish set has 1426 over 40 years old. Therefore, it is possible to analyse how age affects typing. The gender ratio was surprisingly similar in the Finnish and English datasets with 66.8% and 67.5% females, and 30.1% and 27.6% males respectively. 88.4% of the English participants and 48.5% of the Finnish participants used either AC or SB. For more detailed information, see Table 4. Android was the more popular operating system among Finnish participants (67.1%) compared to English (41.6%). iOS was used 32.9% by Finnish and 58.0% by English participants.

The most common typing posture was writing with two thumbs (EN:82.3% FI:81.8%). The thumbs are also the fastest way to type (EN:42.0 WPM, FI:46.9 WPM). The next two most popular postures were the right index finger (EN:2.63%, FI:6.70%) and the right thumb (EN:7.18%, FI:4.86%) which have the typing average of 30.3 WPM in English and 30.5 WPM in Finnish. Younger participants use mostly two thumbs to write. Between ages 15 and 30, 91.6% of Finnish participants and 83.9% of English participants use thumbs. The usage of thumbs gradually decreases in favour of the right index finger; only 23.6% of Finnish 61- to 70-year-olds use thumbs while 49.7% utilize the right index finger. While a similar tendency can be seen in the English dataset, 46.0% of 60- to 70-year-olds still report using thumbs.

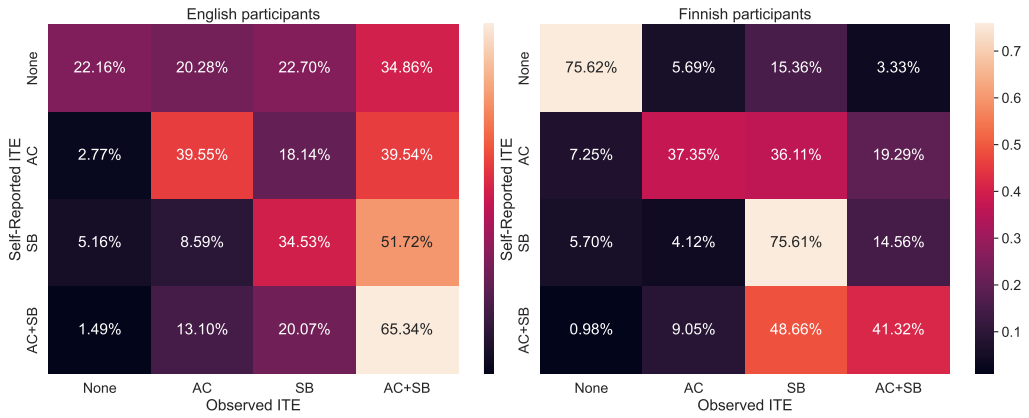
Table 2: Dataset demographics.

	English	Finnish
Participants	24 750	8481
Age	25.9 (10.6)	28.8 (std:11.7)
Gender (f/m)	67.5%/27.6%	66.8%/30.1%
iOS/Android	58.2%/41.8%	34.0%/66.0%
Sentences	1525	550
Avg. WPM	40.2 (std:13.9)	43.6 (std:13.1)
Avg. KSPC	1.09 (std:0.12)	1.13 (std:0.11)
Avg. BSPC	0.061 (std:0.041)	0.068 (std:0.047)

5.2 Self-reported ITE

The challenge in online studies is that the mental state of the participants and the accuracy of the questionnaire answers are uncertain. In the questionnaire, participants are required to report the ITE methods they used during the test. We also observe which ITE methods participants use during the test from the labelled typing data. The difference between self-reported and observed ITE is notable in both languages, but especially among the English participants. The accuracy of self-reported ITE methods among English participants was only 34.6% while for the Finnish participants, it was 67.7%. The confusion matrix in Figure 2 shows the differences between observations and self-reported ITE usage.

The main reason for the low accuracy among English participants is that 79.8% of the participants who self-reported that they did not use ITE methods, used either AC (20%), SB (23%), or both (35%). 40% of the self-reported AC users also used the SB and 53% of the SB users used AC. Finnish participants reported the ITE usage



(a) English (b) Finnish

Fig. 2: Confusion matrix of the self-reported and observed ITE methods.

Table 3: Typing speed average WPM(std) for self-reported and observed ITE methods.

English				
ITE	None	AC	SB	Both
Self-reported	39.4(13.2)	44.5(14.7)	32.8(11.6)	39.3(13.2)
Observed	38.4(13.3)	49.1(14.3)	35.5(12.5)	38.2(12.3)
Finnish				
Self-reported	44.8(13.4)	43.6(11.6)	39.7(12.6)	39.7(11.6)
Observed	45.9(13.0)	40.7(12.3)	41.6(12.8)	41.1(12.6)

more accurately. Still, the participants used the SB more often than the ITE methods they reported. 36% of self-reported AC users used only the SB and 19% both. 49% of Finnish participants who reported that they used both AC and SB, used only SB. The reason why participants reported their ITE usage incorrectly could be because they were uncertain about the methods they used during the test or that they answered which methods they typically use in everyday life instead which can differ from the ones they used during the test.

The average typing speed for both self-reported and observed ITE users are similar as listed in Table 3. The greatest difference in WPM is between AC users. The typing speed of self-reported English AC users was 44.5WPM while for the observed users the typing speed was 4.5WPM higher. On the other hand, the Finnish self-reported AC users typed on average 2.9WPM slower than those who were observed to use AC.

5.3 ITE statistics

The dataset consists of 1525 English sentences and 550 Finnish sentences. The English set has 3378 and Finnish sentences 2008 unique words. 80.8% of the unique English

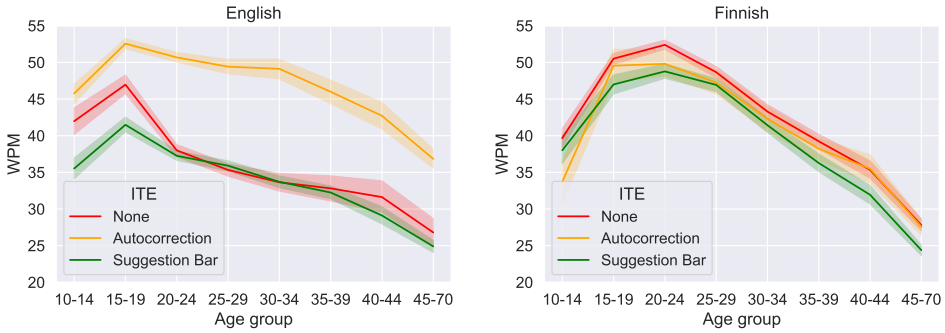
words were autocorrected at least once and 79.3% were selected from the SB at least once. Respectively, 93.5% of the Finnish words were autocorrected and 91.3% selected. The total number of ACs is 229 807 in English and 14 754 in Finnish and the total number of SB selections is 113 389 in English and 20 836 in Finnish.

6 Analysis on Autocorrection and Suggestion Bar

In this section, we use the typing dataset to analyse the effect AC and SB have on typing and the differences between the devices and languages. First, we discuss how the age and operating system direct the usage of ITE methods. Next, we explore how the AC and SB are utilized; how often (utilisation rate), for which words, and how they are utilised in error correction. Then, we show how different devices have different ITE method accuracy and how it affects the typing. Lastly, we shortly investigate the differences between participants' self-reported ITE usage to the usage observed from the data.

6.1 Age

Age is one of the most influential factors affecting the typing performance. The average typing speed decreases as we age due to physical and cognitive changes as can be seen in Figure 3. The results indicate that children between the ages of 10 to 14 are still learning typing and achieve the peak in typing speed around 15 to 19 years after which the speed starts to decelerate. The fastest age group among the English participants is from 15 to 19 years while among the Finnish participants, the best typing age lasts until 24 year-old. However, the typing speed decelerates then faster within Finnish participants than in English.



(a) English (b) Finnish

Fig. 3: Typing speed for each age group. Error bars show the confidence interval. The typing speed decreases after the mid-twenties.

Table 4: Typing speed (WPM) and usage rate of Autocorrection (AC) and Suggestion Bar (SB) by OS.

	None	AC	SB	Both
English				
iPhone	6.2%	37.3%	3.1%	53.3%
WPM	45.8	50.4	41.5	38.7
Android	19.2%	5.00%	49.3%	26.6%
WPM	35.0	35.0	35.0	36.9
Finnish				
iPhone	68.9%	26.7%	-	-
WPM	48.0	40.5	-	-
Android	42.5%	2.57%	42.2%	12.80%
WPM	44.2	41.4	41.4	41.1

6.2 Device

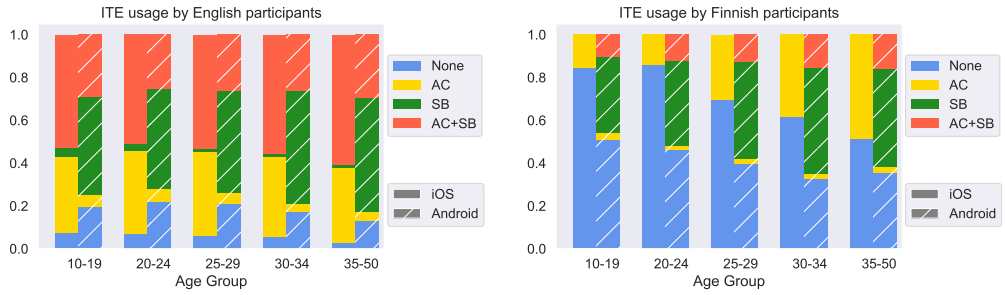
The user optimizes their typing to fit the design of the touch keyboard and the ITE methods. The keyboard and ITE performance depend on the type of the device, especially on the operating system on the device. In the typing dataset, the participants used either iPhone or Android devices, and as is seen in Table 4, the typing speed and the popularity of AC and SB depends on the operating system.

iPhone users are the fastest typing group in both languages. AC aids English participants to achieve the fastest typing speed of 50.4 WPM while Finnish AC users type 40.5 WPM which is 7.5 WPM slower than those who type without any ITE methods. On English iPhone devices, the SB is used along with AC and the users who use both methods type 41.5 WPM. Finnish iPhone devices do not support the SB.

Android users prefer using the SB instead of AC. The participants who used Android had similar average typing speeds whether they used any ITE methods or not. The English Android users type on average 35.0 WPM except if they use both AC and SB in which case the typing speed is 36.9 WPM. The fastest Finnish Android users are again the ones who do not use any ITE methods. The typing speed without any ITE is 44.2 WPM, while with ITE, participants type with 41.4 WPM.

AC is extremely popular with English iPhone users with 90.6% of users using it during the test, while only 31.6% of English Android users used AC. AC is less popular with Finnish participants as only 26.7% of iPhones and Android users 15.37% use it. While the usage of the SB in English results in the slowest typing speed, it is a popular tool, as 66.0% of English participants and 40.4% of Finnish participants use the SB at least once during the test. There is a clear difference between the preferences of Android and iPhone users. iPhone users use mainly AC and SB is used alongside the AC and rarely alone. On the Android devices, it goes the other way around, SB is the more popular among Android users. 75.9% of English and 55.0% of Finnish Android users use the SB while with English iPhone users the amount is 56.4%.

AC and SB are used more among the older participants. The increase is shown in Figure 4 for each operating system. Especially the use of the Suggestions increases with the users who have them available. The increase is more noticeable with Finnish than English participants because ITE techniques are popular with English participants in all age groups. Figure 3 shows that AC in English makes the decline in speed due to age less steep.



(a) English (b) Finnish

Fig. 4: ITE usage rate by different age groups. ITE usage increases with age.

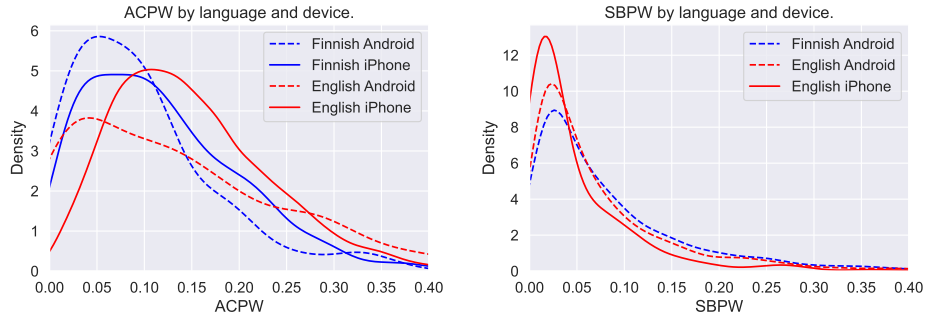
6.3 Utilisation Rate

The frequency of ITE usage varies, some participants use a certain method constantly, and some less often. The utilisation rate measures how often ITE method is used by dividing the number of uses by the total number of words in the sentence. To measure the utilisation rate of the AC and SB, we introduce the ITE per word (ITEPW) metric. It measures the percentage of the words the user typed during the test that was completed by selecting a word from the Suggestion Bar (SBPW) or corrected with Autocorrection (ACPW). The Kernel Density Estimation [36] of ITEPW for both languages and Operating systems is plotted in Figure 5. Next, we show how the utilisation rate affects typing speed.

Autocorrection

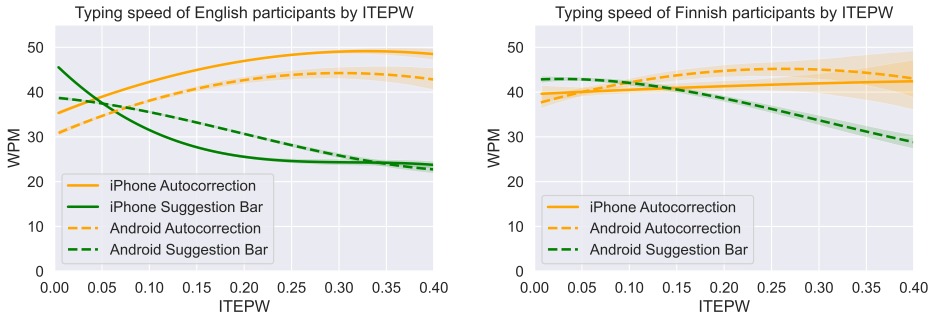
The English participants who used AC on iPhone devices, trigger AC more often than other participants. Figure 5a shows the AC utilisation rate (ACPW) distribution for iPhone and Android devices in both languages. The median of ACPW for English iPhone AC users is 13.7% while for Android the median is 11.7%. Compared to Android, very few iPhone users have ACPW less than 5% meaning that almost all of the users use AC regularly. Similar differences between devices can be observed among Finnish participants as the median for iPhone users is 10.6% and for Android 8.1%. The usage rate is lower among Finnish participants.

The typing is a trade-off between typing accuracy and speed. The faster the user types, the more typing mistakes they make. Therefore, when the typing speed



(a) Autocorrection (b) Suggestion Bar

Fig. 5: ITEPW density distribution for iPhone and Android users. AC is used more often than SB and is in general more activily used on iPhones.



(a) English (b) Finnish

Fig. 6: Effect of ITE utilisation rate (ITEPW) to typing speed. A high utilisation rate of AC increases typing speed, but a high SB usage rate decreases the typing speed.

increases, more AC is triggered because the user makes more typing errors. We illustrate this phenomenon in Figure 6 where the polynomial model (order=2, ci=0.95) is fitted to the data to visualize the relationship between the typing speed and the utilisation rate (ITEPW). According to the model, the typing speed increases with the utilisation rate until 30% ACPW in English and 25% ACPW in Finnish after which the typing speed saturates. In English, the shape of the fitted line is the same regardless of the device, however, the average speed is lower on Android devices. However, in Finnish, the AC on Android devices seems to improve more with the increased use compared to iPhone devices, which could indicate that the AC performance is better on Android devices.

Suggestion Bar

The average utilisation rate is lower for SB than for AC. Most users select suggestions for 2% to 5% of the words as is seen in the distribution of SBPW in Figure 5b. Android users utilise SB more often than iPhone users. The median of SBPW for English Android users is 4.5% while for iPhone users 2.4%. SB is not available on the Finnish iPhone keyboard, however, Android users have a utilisation rate of 5.5%.

From the polynomial model (order=3, ci=0.95) fitted to the data in Figure 6, we can see that, unlike AC, the SB is the most beneficial for typing speed when it is used to fill less than 10% of the words. The more often Suggestions are selected during the typing, the slower the typing becomes. In English, especially on iPhone devices, the drop in typing speed is faster when the utilisation rate increases compared to Finnish. The Finnish users benefit more from higher SBPW.

6.4 Word Length and Keystrokes

The length of the words affects the typing speed. The longer the word the slower the user types, even if word length is normalised by the number of characters. The average length of the word in English test sentences is 4.23 characters, while the number is 6.57 in Finnish. The effect of the word length is modelled with linear regression which is plotted in Figure 7, where the blue line shows how the typing speed decreases when the average word length in the sentence increases and no AC or SB is used. The Finnish participants are less affected by the word length than the English participants. The typing speed decreases similarly even if AC is used.

Table 5: The average word length of the sentences and word length average of Autocorrected and Suggestion Bar selected words. Standard deviation is inside the brackets.

English			
	All words	Autocorrection	Suggestion Bar
Word length	4.23 (2.27)	4.86 (2.40)	6.95 (2.27)
Finnish			
Word length	6.57 (3.38)	7.47 (3.38)	8.50 (3.23)

When the SB is utilised, the average word length of the sentence affects less to the typing speed. In Figure 7, we visualise the relationship between the word length and typing speed by fitting a linear regression model to the data of the participants who use the SB. The blue line in Figure models all sentences typed by the SB users, the green line is the sentences where SB was used without errors and the orange line sentences with SB errors. Among the English participants, when SB is used, the typing speed is the same regardless of the word length. When the average length of the words is longer than 7, users type faster by utilising the SB In Finnish, the longer words slow typing even when SB is used, however longer the word, the less differences

there are in the typing speed. SB is best used for longer words and during the test, the participants used it for the long words. The average word length for words which were completed by the SB was 6.95 characters in English which is 2.72 characters longer than the average of the test sentences. In Finnish, the average word length of selected words was 8.50 characters while the overall average was 6.57 characters. The importance of the accuracy of the suggested words is also seen in Figure (orange line). Selecting incorrect words to complete the words is slow.

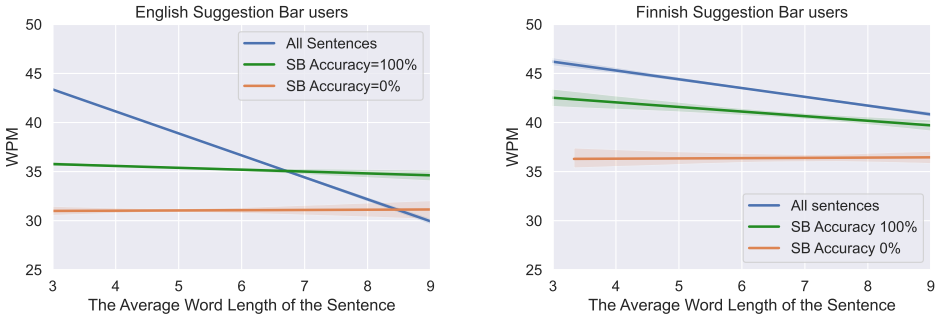
Table 6: The average of number of keystrokes (KSPC) and backspaces (BSPC) for different user groups. Standard deviation is inside the brackets.

	None	Autocorrection	Suggestion Bar
English			
KSPC	1.14 (0.10)	1.12 (0.084)	1.08 (0.10)
BSPC	0.071 (0.050)	0.063 (0.040)	0.057 (0.039)
Finnish			
KSPC	1.14 (0.10)	1.15 (0.098)	1.13 (0.095)
BSPC	0.070 (0.051)	0.078 (0.049)	0.065 (0.039)

Regardless of the language, for the participants who type without any ITE methods, the number of keystrokes and backspaces are similar. As is listed in Table 6, participants without ITE use 1.14 Keystrokes per Character (KSPC) and 0.07 Backspaces per Character (BSPC). The utilisation rate of the AC correlates positively with KSPC (en: r=0.28, fi: r=0.36, p¡0.0001) and BSPC (en: r=0.26, fi: r=0.35 p¡0.0001). AC increases the number of keystrokes and backspaces slightly because correcting AC errors requires more typing. Therefore lower accuracy of AC causes more keystrokes which are seen as the correlation between the number of keystrokes (en: r=0.48, fi: r=0.53, p¡0.0001) and backspaces (en: r=0.53, fi: r=0.54, p¡0.0001) and the number of corrected AC errors. While the SB users type more slowly, they use the least KSPC of all participants. The more often the SB is used, the fewer keystrokes are necessary. The correlation between the SB utilisation rate and the KSPC is negative (en: r=0.55, fi: r=0.37, p¡0.0001). The utilisation of the SB does not affect the number of backspaces much, most likely, because users tend to select incorrect suggestions and use backspaces to correct them.

6.5 Error Correction

Mobile keyboard lacks tactile feedback and is smaller than physical keyboards which causes users to make more mistakes during typing and correcting errors is slower. ITE methods help with the challenges of mobile keyboards by providing error correction and ways to avoid errors. Both AC and SB are used to correct user's typing errors. AC corrects typing mistakes automatically, while SB requires the selection of the word. The number of the typing errors is calculated by edit distance. The edit distance is calculated between the sentence user typed and the original sentence they were



(a) Finnish (b) English

Fig. 7: The average WPM by the average word length of the sentence. Blue line includes all sentences by SB users. Orange line all sentences with the SB accuracy of 100% and green line sentences with the accuracy of 0%. The average word length affects less on the typing speed when SB is used.

supposed to type. AC corrects on average on English iPhones 1.58 characters during correction and on Android devices 2.1 characters. For Finnish users, the number of characters corrected during autocorrection was smaller with 1.45 characters for iPhone and 1.47 characters per word for Android users.

We assume that the SB is used for error correction when the length of the word is identical to the suggested word. English participants who use Android devices used 14.25 % of the SB selections to correct on average 1.8 characters. On iPhone, only 3.13% of the selected words were used for correction while 2.02 characters were replaced on average. The error correction with SB is more popular in Finnish. Finnish participants use 33.13% of the suggestions to correct typing errors. The number of errors corrected was on average 1.81 characters in a selected word. SB can also correct errors when it fills the word for the user. In English 11.45% of the selected words were used to correct errors while filling in the rest of the word. 1.68 characters were corrected. Similarly, in Finnish 26.15% of the SB selected words were used to correct while completing the word. A little less typing errors were corrected during the the word completion compared to when SB was only used for error correction. In English 1.68 characters and in Finnish 1.40 characters were corrected. The corrected words were shorter than the SB word length average. The length of the corrected words in English was 5.43 (std2.43) and in Finnish 7.02 (std:3.03) characters.

6.6 Accuracy

The accuracy of the AC and SB affects significantly the typing experience because the accuracy determines the correctness of the autocorrections and the quality of the provided suggestions. ITE methods are built on language models, therefore, the accuracy depends on the language, the used device and the context. In the case of AC, low accuracy causes incorrect correction by replacing the typed word with a wrong one. The accuracy of the SB represents the quality of the suggestion provided for the

user. The user selects a word from a list of words thus the wrong selections are rarely selected without intention. Because the data does not contain the options the user is given, we are only able to observe the selected words.

Autocorrection

Unlike the SB users, AC users cannot select when AC is used thus errors cannot be avoided. Correcting AC errors slows the typing speed of the user. When we observe only sentences where AC is used, English AC users type with AC on average 51.58 WPM, however, in case of an AC error, the typing speed drops to 40.11 WPM. Similarly in Finnish, AC users type on average a sentence without AC errors at a speed of 45.21 WPM but with errors, the speed decreases to 35.98 WPM.

AC is used actively on iPhone devices as almost all English participants use it. The utilisation rate is also higher on iPhone devices compared to Android, see section 6.3. On the other hand, the accuracy of AC on iPhone is also lower (error rate=19.4%) which indicates that the threshold for AC is lower. As is listed in Table 7, the error rate is 11.0% on Android devices. The threshold determines when AC should be triggered and it is a trade-off between the accuracy of the model and the number of typing errors corrected. A lower threshold causes AC to be more aggressive. Finnish AC is less accurate than English: 23.0% error rate on iPhone and 21.8% error rate on Android. The differences in AC accuracy between languages can be observed by comparing the number of incorrect letters in AC errors. In English, the edit distance between the correct and incorrect AC is 2.01 characters while in Finnish edit distance is 2.92 characters.

The user has to notice and correct the AC errors themselves. However, ACs can be unexpected and errors even more so. AC can change the correctly typed word to the wrong one. The probability of incorrectly correcting a word is lower on iPhone devices (English: 9.2% and Finnish: 10.6%) compared to Android (English: 10.4% and Finnish: 12.6%). Users might not immediately notice that the word has changed. English participants corrected 88.17% of the AC errors while Finnish users corrected 95.00% of the errors.

Table 7: The Error Rate of the Autocorrection and Suggestion Bar.

ITE Error Rate		
English		
OS	Autocorrection	Suggestion Bar
Android	11.9%	11.2%
iPhone	19.4%	16.0%
Finnish		
Android	21.8%	15.1%
iPhone	23.0%	-

Suggestion Bar

While one would think that users only select correct words from the SB, in fact, 15.9% of the Finnish and 13.3% of the English selected words were incorrect. Users pick the incorrect suggestion because the correct one is not available. Instead, a similar word is selected and corrected. Almost all of the incorrect suggestions were corrected. Finnish participants corrected 98.00% and English participants 97.15% of the selection errors. The similarity of the incorrect selection and the target word can be observed by comparing the number of letters the selection adds to the number of letters that need correcting. The incorrect selection in Finnish was on average longer than the SB selection which was correct. For Finnish the incorrect words were on average 10.44 characters while the average on all SB words was 8.3 characters. In English, the length difference was not significant: The word length of incorrect words is 7.03 and for all selected words 6.95 characters. The number of added letters on average is 3.48(std:2.02) for English and 3.61(std:2.24) for Finnish while the edit distance (the number of edits to correct the word) is 2.06(std:1.72) characters for English and 3.01(std:2.27) characters for Finnish. Especially for English, not many edits are needed to correct the word.

7 Discussion

We collected and processed a large-scale typing dataset in English and Finnish languages to provide a dataset to analyse the performance and effect of AC and SB on iPhone and Android devices. To our knowledge, the dataset presented in this paper is the only set available that contains typing data gathered with the user's device and keyboard with robust and detailed, keystroke-level ITE-related labels. The large dataset enables robust estimates for analysis and provides typing data on users of different ages and backgrounds. The data was gathered with a large variety of iPhone and Android devices with different operating system versions. User's keystrokes, AC and SB uses along with other device-produced auto-inputs are labelled on the dataset at a keystroke level. The ITE utilisation rate, the error rate and the correction rate of the ITE methods were also added to sentence and participant-level data. In addition, the dataset includes a list of all autocorrected and user-selected suggestion words along with the words before the correction or suggestion and the original word.

While the typing dataset is a large-scale dataset with participants of versatile backgrounds and devices, the typing can only be observed from the changes in the text input field. The missing information on what is happening on the graphical interface is crucial for certain types of analysis, such as which SB suggestions the user was provided, and AC has also feature where the predicted word is shown above the typed word. The data itself is noisy due to all the different devices and operation system versions thus there are always exceptions in keystroke logging which are not accounted for. However, the language model and the user adapt to each other over a long time, and therefore, it is important to observe typing with the user's device. The labels are also limited only to the AC and SB as the Gesture recognition was difficult to distinguish from the Suggestions. As is often the case with online studies, the experiment is

less controlled, meaning we are not able to tell for certain if the test is performed correctly or if the questionnaire is answered accurately. The importance of labelling the data from the observed information was discovered when the accuracy of self-reported ITE usage among English participants was only 35%. The problem of the self-reported answers should be taken into consideration in future studies.

The challenge to evaluate the performance of ITE methods and the language models is that they can only be observed from outside and there are many variables affecting the results. We found that the iPhone and Android users utilize two different typing styles which indicates that the design of the keyboard and the language model affect the user's choice of typing strategy. Over 90% of iPhone users utilise AC. AC on iPhone devices has a lower triggering threshold compared to Android devices, and therefore, assumed errors are corrected more often which causes a higher error rate. However, it seems that the lower threshold serves the users well as the iPhone users are the fastest typing user group. AC users balance between the speed and error trade-off: The user increases the typing speed to the point where AC is still able to correct most of the errors. However, the active use of AC results in a higher rate of uncorrected errors in the text. Even though the use of SB is more popular among the older participants compared to participants who use Android, SB is used in moderation on iPhone devices. On the Finnish iPhone keyboard, the SB is not even available Android users, on the other hand, lean towards using the SB to correct and avoid typing errors which results in slower typing speed. AC is rarely used by itself on Android devices.

The accuracy of the Language models depends on the language. Finnish has lower accuracy on both ITE methods. The ITE utilisation rate is also lower compared to the rate of the English participants who almost all use some sort of ITE to aid their typing. In addition to the higher error rate, the number of letters to be corrected is also higher in Finnish and, therefore, the penalty for correcting AC errors is even higher. The cost of error in AC is high. The low accuracy on AC causes a drop of 10 WPM in the average typing speed. When the accuracy is low enough, typing without AC is more beneficial. Instead of AC, Finnish participants preferred using SB to correct typing errors as 59.28% of the time suggestions corrected errors, for English the rate was only 25.7%. The SB performs better than AC even with low accuracy as the user has control when it is utilized. The SB supports a fast typing pace the best when it is selectively used only for difficult or long words.

To our knowledge, the difference between iPhone and Android devices has not been studied before. Previously it was known that Android users type slower [9] but the reasons have not been explored. AC and SB have been studied in laboratory environments and with computational models previously. While they provide controlled information on typing, they do not account for the performance of ITE methods which depends on the used device. While a large-scale typing dataset is more noisy and dependencies are more difficult to discern, the data shows the performance of the real language models developed for iPhone and Android devices. In this paper, we were able to demonstrate how the two operating systems and the used language steer users to different typing strategies.

8 Conclusion

We collected and labelled a large-scale dataset to study how AC and SB with real devices perform in real life and how the performance affects typing. We found that there is a significant dissimilarity between the typing strategy of the iPhone and Android users, which might be due to the design decisions of the touch keyboard and ITE methods on the devices. Almost all iPhone users use AC which steers faster typing while increasing the typing errors which are then corrected by relatively aggressive AC. Android users prefer SB to avoid typing errors. Finnish participants who suffer lower accuracy on ITE methods also prefer using SB as an error correction method instead of AC. In general, Finnish participants utilised ITE methods less often compared to English participants who almost all used some combination of ITE methods.

In this paper, we demonstrated that the typing dataset can be used to analyse the performance of ITE methods and how they affect the user. While we observed that there is a significant difference between iPhone and Android devices, the ITE and touch keyboard design decisions which lead to steering users to certain typing strategies are left for future work. Languages are a key factor in the differences in language models, and therefore, the linguistic aspect of the ITE method performance should be studied more closely in the future. The dataset contained two languages: English and Finnish, however, it is easy to expand to other languages organising the same user study as all the relevant codes are publicly available and support multiple languages. The dataset contains over 50 000 participants with various backgrounds and devices. A large dataset enables studying certain user groups, for example, non-natives or participants over 40 years old. We found that inaccuracy in the self-reported answers compared to observation from the data. In the future work, we explore the differences in more detail. As mobile typing is becoming one of the most important ways to produce text, it is important to study how the users type on their devices and which key factors have the most impact on the typing experience.

Acknowledgments. We would like to thank all volunteers who participated in typing test. This work was supported by Emil Aaltonen Foundation, KAUTE Foundation, Foundation for Aalto University Science and Technology and Google. The computational resources were provided by Aalto ScienceIT.

Declarations

Conflict of interest. The authors have no relevant financial or non-financial interests to disclose.

References

[1] Whitelaw, C., Hutchinson, B., Chung, G., Ellis, G.: Using the web for language independent spellchecking and autocorrection. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 890–899 (2009)

[2] Ghosh, S., Kristensson, P.O.: Neural networks for text correction and completion in keyboard decoding. arXiv preprint arXiv:1709.06429 (2017)

[3] Bellegarda, J.R.: Pragmatic pertinence: A learnable confidence metric to assess the subjective quality of lm-generated text (2023). ISCA

[4] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)

[5] Banovic, N., Sethapakdi, T., Hari, Y., Dey, A.K., Mankoff, J.: 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, pp. 1–12 (2019)

[6] Jokinen, J.P., Sarcar, S., Oulasvirta, A., Silpasuwanchai, C., Wang, Z., Ren, X.: Modelling learning of new keyboard layouts. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 4203–4215 (2017)

[7] Banovic, N., Rao, V., Saravanan, A., Dey, A.K., Mankoff, J.: Quantifying aversion to costly typing errors in expert mobile text entry. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 4229–4241 (2017)

[8] Klein, J., Moon, Y., Picard, R.W.: This computer responds to user frustration. In: CHI'99 Extended Abstracts on Human Factors in Computing Systems, pp. 242–243 (1999)

[9] Palin, K., Feit, A.M., Kim, S., Kristensson, P.O., Oulasvirta, A.: How do people type on mobile devices? observations from a study with 37,000 volunteers. In: Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1–12 (2019)

[10] Holz, C., Baudisch, P.: Understanding touch. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2501–2510 (2011)

[11] Feit, A.M., Weir, D., Oulasvirta, A.: How we type: Movement strategies and performance in everyday typing. In: Proceedings of the 2016 Chi Conference on

Human Factors in Computing Systems, pp. 4262–4273 (2016)

[12] Varcholik, P.D., LaViola Jr, J.J., Hughes, C.E.: Establishing a baseline for text entry for a multi-touch virtual keyboard. International Journal of Human-Computer Studies **70**(10), 657–672 (2012)

[13] Jiang, X., Li, Y., Jokinen, J.P., Hirvola, V.B., Oulasvirta, A., Ren, X.: How we type: Eye and finger movement strategies in mobile typing. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2020)

[14] Sarcar, S., Jokinen, J.P., Oulasvirta, A., Wang, Z., Silpasuwanchai, C., Ren, X.: Ability-based optimization of touchscreen interactions. IEEE Pervasive Computing **17**(1), 15–26 (2018)

[15] Leiva, L.A., Kim, S., Cui, W., Bi, X., Oulasvirta, A.: How we swipe: A large-scale shape-writing dataset and empirical findings. In: Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction, pp. 1–13 (2021)

[16] Kukich, K.: Techniques for automatically correcting words in text. ACM computing surveys (CSUR) **24**(4), 377–439 (1992)

[17] Bassil, Y., Alwani, M.: Post-editing error correction algorithm for speech recognition using bing spelling suggestion. arXiv preprint arXiv:1203.5255 (2012)

[18] Banovic, N., Grossman, T., Fitzmaurice, G.: The effect of time-based cost of error in target-directed pointing tasks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1373–1382 (2013)

[19] Alharbi, O., Stuerzlinger, W., Putze, F.: The effects of predictive features of mobile keyboards on text entry speed and errors. Proceedings of the ACM on Human-Computer Interaction **4**(ISS), 1–16 (2020)

[20] Alharbi, O., Stuerzlinger, W.: Auto-cucumber: The impact of autocorrection failures on users' frustration. In: Graphics Interface 2022 (2021)

[21] Weir, D., Pohl, H., Rogers, S., Vertanen, K., Kristensson, P.O.: Uncertain text entry on mobile devices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2307–2316 (2014)

[22] Arnold, K.C., Chauncey, K., Gajos, K.Z.: Predictive text encourages predictable writing. In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 128–138 (2020)

[23] Buschek, D., Zürn, M., Eiband, M.: The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In: Proceedings of the 2021 CHI Conference on Human Factors

in Computing Systems, pp. 1–13 (2021)

[24] Quinn, P., Zhai, S.: A cost-benefit study of text entry suggestion interaction. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 83–88 (2016)

[25] Dunlop, M.D., Crossan, A.: Predictive text entry methods for mobile phones. Personal Technologies **4**, 134–143 (2000)

[26] Caine, K.: Local standards for sample size at chi. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 981–992 (2016)

[27] Vertanen, K., Kristensson, P.O.: A dataset of noisy typing on qwerty keyboards. In: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces. IUI '23 Companion, pp. 251–254. ACM, ??? (2023). https://doi.org/10.1145/3581754.3584174

[28] Dhakal, V., Feit, A.M., Kristensson, P.O., Oulasvirta, A.: Observations on typing from 136 million keystrokes. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2018)

[29] Yleisradio: Ylen suomenkielisen uutisarkiston selkouutiset 2011-2018, sekoitettu, Korp. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2019121204

[30] Aller Media Oy: Suomi24 virkkeet -korpus 2001-2017, Korp-versio 1.2. Kielipankki (2019). http://urn.fi/urn:nbn:fi:lb-2020021803

[31] Kristensson, P.O., Vertanen, K.: Performance comparisons of phrase sets and presentation styles for text entry evaluations. In: Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, pp. 29–32 (2012)

[32] Vertanen, K., Kristensson, P.O.: 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, pp. 295–298 (2011)

[33] Graff, D., Cieri, C.: English gigaword LDC2003T05 (2003) (2003). https://catalog.ldc.upenn.edu/LDC2003T05

[34] Franco-Salvador, M., Leiva, L.A.: Multilingual phrase sampling for text entry evaluations. International Journal of Human-Computer Studies **113**, 15–31 (2018)

[35] Wobbrock, J.O.: Measures of text entry performance. Text entry systems: Mobility, accessibility, universality, 47–74 (2007)

[36] Chen, Y.-C.: A tutorial on kernel density estimation and recent advances. Biostatistics & Epidemiology **1**(1), 161–187 (2017)