Logging Keystrokes in Writing by English Learners

Georgios Velentzas[1], Andrew Caines[2], Rita Borgo[1], Erin Pacquetet[3], Clive Hamilton[4], Taylor Arnold[5], Diane Nicholls[6], Paula Buttery[2], Thomas Gaillat[7], Helen Yannakoudakis[1], and Nicolas Ballier[8]

[1] King's College London, U.K., [2] ALTA Institute & Computer Laboratory, University of Cambridge, U.K.
[3] University at Buffalo, U.S.A., [4] Université Paris Cité, CLILLAC-ARP, France
[5] University of Richmond, U.S.A., [6] English Language iTutoring (ELiT), U.K.
[7] Université de Rennes, LIDILE, France, [8] Université Paris Cité, CLILLAC-ARP & LLF, France

geovelentzas@gmail.com, {rita.borgo, helen.yannakoudakis}@kcl.ac.uk
{andrew.caines, paula.buttery}@cl.cam.ac.uk, diane.nicholls@englishlanguageitutoring.com
tarnold2@richmond.edu, thomas.gaillat@univ-rennes2.fr
{clive.hamilton, nicolas.ballier}@u-paris.fr

Abstract

Essay writing is a skill commonly taught and practised in schools. The ability to write a fluent and persuasive essay is often a major component of formal assessment. In natural language processing and education technology we may work with essays in their final form, for example to carry out automated assessment or grammatical error correction. In this work we collect and analyse data representing the essay writing process from start to finish, by recording every key stroke from multiple writers participating in our study. We describe our data collection methodology, the characteristics of the resulting dataset, and the assignment of proficiency levels to the texts. We discuss the ways the keystroke data can be used – for instance seeking to identify patterns in the keystrokes which might act as features in automated assessment or may enable further advancements in writing assistance – and the writing support technology which could be built with such information, if we can detect when writers are struggling to compose a section of their essay and offer appropriate intervention. We frame this work in the context of English language learning, but we note that keystroke logging is relevant more broadly to text authoring scenarios as well as cognitive or linguistic analyses of the writing process.

Keywords: keystroke logging, language learning, essay writing

1. Introduction

Given that for many the authoring of documents involves typing on a personal computer, a great deal of information about the writing process can potentially be captured from keystroke data. Relating to education technology for language learners, we can potentially detect when learners are struggling with their writing, which might enable supportive interventions to aid the learner. The same is true for text authors more generally. There are also implications for cognitive science more widely, in that keystroke data can give us insights into linguistic creativity and aspects of language complexity in production. It may be that we can use keystroke data to detect bursty events or copy-and-paste actions which might point to the use of generative AI for text generation, and malpractice in education or assessment settings.

We have compiled a dataset of texts, keystroke logs and metadata for public release. It contains a copied text and creative essay written in English by 1,006 crowdsourced participants, both native speakers and non-native speakers of the language. Approximately one-fifth of the participants self-identified as native speakers of English. Within the non-native speaker group, a total of 42 first languages are represented amongst the participants: Polish, Portuguese and Spanish being the most common. The median age of our participants is 26, with a minimum of 18 and maximum of 75 years. The essays they wrote have been assigned a proficiency level within the Common European Framework of Reference for Languages[1]. This information will enable analysis of the keystroke data in search of patterns which correlate with writing proficiency.

In this paper we describe our method for preparing the text authoring interface and the JavaScript keylogger, our procedure for participant recruitment and data collection, and our preparation for dataset release including the grading of essays by multiple assessors in order to arrive at proficiency levels for each one. We report on our analyses of the features and statistics found in the resulting dataset, correlations with proficiency level, and ways in which we think this dataset enables future research into the writing process. In particular we believe that the dataset can be used to associate keystroke patterns and dynamics with

[1] http://www.coe.int/lang-cefr

LREC-COLING 2024, pages 10725–10746

proficiency levels in research around automated essay assessment. We also propose that by detecting when writers are struggling with linguistic constructions we can offer supportive interventions which aid in text completion and language learning. The dataset also offers the possibility of cognitive and linguistic analyses into the ways that texts are constructed and the smaller structures which underpin sentences and paragraphs. The **K**ing's College London & **U**niversité **Pa**ris Cité Keys (**KUPA-KEYS**) dataset is publicly available for non-commercial use[2] and our research code for data collection is open-source as detailed in Section 3.

2. Related Work

Previous studies on keystroke data have been wide-ranging, at times analysing individual variation for stylometric or biometric purposes (Giot et al., 2009; Roffo et al., 2014; Plank, 2018; Udandarao et al., 2020). Such was the success of individual identification, documented in numerous papers (Tappert et al., 2010; Stewart et al., 2011; Monaco et al., 2012, 2013; Kang and Cho, 2015), that it has been used for security and authentication purposes – though in turn this brought risks of adversarial attacks and impersonation which had to be addressed (Monaco and Tappert, 2016). The biometric work using keystroke data for security reasons has continued into the deep learning era of neural network machine learning (Acien et al., 2020; Maiorana et al., 2021; Stragapede et al., 2023).

Other applications of keystroke analysis include cognitive science research into pausing during writing, and the underlying reasons for it (Galbraith and Baaijen, 2019), as well as investigations intro translation processes (Schaeffer et al., 2016) and the causes of spelling errors (Baba and Suzuki, 2012). Keystroke data has also been used as additional features for improving natural language processing tools (Goodkind and Rosenberg, 2015; Plank, 2016), and has been used in linguistic analyses of hierarchical structures emerging as written texts take shape (Ballier et al., 2019; Leijten et al., 2019; Mahlow et al., 2022).

Of most relevance to our work are previous keystroke datasets, including password studies (Giot et al., 2009) and longer texts (Tappert et al., 2010; Monaco et al., 2012); as well as research into the relationship between keystroke behaviour and writing quality (Zhang et al., 2016), the use of features derived from keystroke data in automated writing evaluation (Chukharev-Hudilainen, 2019; Chukharev-Hudilainen et al., 2019), and in-application support based on the detection of writing errors (Mahlow, 2015).

Datasets generated from participants transcribing a fixed-text are more common, (Allen, 2010, Teh et al., 2011, Feit et al., 2016, Fierrez et al., 2010, Idrus et al., 2013, Giot et al., 2012) with the largest being reported in Dhakal et al. (2018), consisting of 136 million keystrokes produced by 168,000 users. On the contrary, datasets that include free-text are not as extensive and usually incorporate additional data from a transcription task. One of the earliest datasets is Clarkson I (Vural et al., 2014), comprising a total of 840,000 keystrokes produced by 39 participants answering survey questions, transcribing a text, and creating passwords.

Similarly, Banerjee et al. (2014) produced a dataset where participants responded to business-related questions either truthfully or deceptively. Monaco et al. (2015) collected data from 64 undergraduate students typing with both hands, left hand only, and right hand only, while answering computer science-related questions during an exam. Datasets in languages other than English are also available, with the work of Gunetti and Picardi (2005) being in Italian and Montalvão Filho and Freire (2006) including free-text data in Portuguese.

One of the largest datasets available is described in Sun et al. (2016), known as the Buffalo dataset, comprising 2.14 million keystrokes from 148 participants. The authors provided keystroke and mouse movement data across three sessions per subject. Each session included a) transcription of a speech, b) answers to query questions and descriptions of a picture, and c) composing an email and Internet surfing. Lastly, Murphy et al. (2017) provides the Clarkson II dataset, where keystroke data were collected in an uncontrolled environment from 103 subjects who used their computers normally over 2.5 years, yielding a total of 12.6 million keystrokes. Our dataset is distinct from the ones above because we focus on essay writing by both learners and native speakers of English, with both text-copy and free-text composition. In addition we have had the free-text essays graded by trained but non-operational assessors of English, and make the dataset publicly available for research use.

A new Kaggle shared task was launched in October 2023 that involved predicting overall writing quality based on keystroke data[3]. The dataset involved is a large collection of U.S. SATS essays (captured from 5000 participants), scored on

[2]Available from https://huggingface.co/datasets/ALTACambridge/KUPA-KEYS under a Creative Commons Attribution Non-Commercial Share-Alike 4.0 International licence.

[3]"Linking Writing Processes to Writing Quality".

a scale of 0-6, with all alphanumeric characters anonymised. Even though our dataset is smaller, the essays are scored on a scale of 0-12, as explained below, and the alphanumeric characters are preserved. Furthermore, we have included additional keystroke log data obtained from a transcription task. This data may act as a user-specific baseline, allowing for the creation of calibration features; features that can be utilized during the model-building process to cater to personalized applications. It is crucial to note that each individual possesses a unique typing pattern (Leggett et al., 1991; Peacock et al., 2004; Karnan et al., 2011), therefore we should attempt to capture these differences for enhancing model effectiveness and precision.

3. Text Authoring Interface

In order to collect user-generated texts and associated keystroke data, we built our own custom GitHub Pages site[4] with text editing functionality and JavaScript keylogging plug-in[5]. The choice of programming language is important. JavaScript is the *de facto* language of the web and can be run by any standard web browser. This is useful because it means that our software can run directly on any user's machine through a browser without any installation. Also, because the web standards are more stable and open than internal APIs built to capture keylogging in proprietary software such as Microsoft Word, the tool is both relatively stable and easy to adapt to new use cases.

Another benefit of JavaScript is that it has excellent support for processing user inputs such as keypresses and mouse movement. The language is designed around the principle of asynchronous computation, meaning that the timestamp data should remain accurate even on a slower machine or during a period of fast typing. Time records in JavaScript are recorded in milliseconds and should be accurate to within ± 5 milliseconds (Mozilla Foundation, 2023).

The data capture interface stores each submission data in JSON format, consisting of an array of objects where each object represents an event. To create a more informative dataset, we include both user-related and system-related events. User-related events encompass raw information about key down, key up, and mouse clicks. System-related events, on the other hand, comprise details like the actual modifications made to the content, capturing which characters were added or removed, as well as periodic captures of the full text in the text box. More details are provided in the Dataset Description section and there is a screenshot of the authoring interface in the Appendix.

4. Data Collection

We recruited more than one thousand participants using the Prolific crowdsourcing platform[6]. Participants were first of all directed to a Qualtrics survey in order for us to obtain some useful metadata, informed consent and provide task instructions. The metadata we asked for included age, country of residence, native language, keyboard layout, hours per day spent on a computer, years learning English, daily exposure to English, other languages known, level of English and level of education. We only allowed participants who were using either a desktop or a laptop in our study (as opposed to mobile devices) – which was initially filtered by Prolific, but also the keylogger captures additional device information from the user such as operating system and browser version.

After the survey, participants were redirected from Qualtrics to our text authoring site. Participants were required to complete two writing tasks: a **copy-text** task (task 1) & an **essay-writing** task (task 2). The copy task involved re-writing a provided text, a 300-word excerpt from a Steve Jobs speech which was chosen as it contains a high number of distinct English digraphs (197). Additionally, this excerpt is employed as one of the tasks in Sun et al. (2016) – although our dataset serves a different purpose, it has the potential to contribute to the improvement of datasets for user authentication applications.

The essays were written in response to a random selection from a set of 10 'just for fun' prompts from the English learning platform Write&Improve[7]. The just for fun prompts were chosen, as opposed to level-specific prompts, as they are deliberately creative, suitable across different proficiency levels, and tend not to elicit personal information. This is as opposed to prompts which are targeted at beginners of English, for instance, or on topics which entail students writing about daily routines, hometowns or family structures. Two examples of the ten just for fun prompts are given below:

- *A Special Place*. If you could be anywhere in the world right now, where would you choose to be? Describe the place. Why do you want to be there?
- *Unforgettable*. Write a short story with the title 'Unforgettable'. Your story must have a begin-

[4]Source code: https://github.com/CambridgeALTA/keylog-pages

[5]The JavaScript keylogger is available in an open-source GitHub repository: https://github.com/statsmaths/keylog

[6]https://www.prolific.com/

[7]https://writeandimprove.com

ning, a middle and an end. The end must be surprising.

After a pilot study, we found that these two tasks took approximately 30 minutes to complete on average. We paid participants £7.5 GBP / $9.2 USD for answering the metadata-related questions and completing the tasks satisfactorily. In the end after collecting all data, we found that the participants had spent an average of 33.9 minutes for completing the survey, allocating 9.6 minutes to the copy-text task and 13.4 minutes to the essay-writing task.

Our first requirement for data approval was that the submitted essays should be at least 80% of the minimum stated length: initially 250 words but later 150 words, for reasons explained below. We reviewed participants' data in order to detect and reject texts which had obviously been copy-pasted from external sources, or generated by pre-trained large language models. The former was addressed by identifying anomalous, bursty keystroke patterns, with ongoing efforts being made to handle the latter.

For instance, we utilized open-source generative-AI text classifiers, such as the tool released by OpenAI on January 31st, 2023. However, this tool was discontinued six months subsequent to its release due to its low accuracy. Consequently we have opted not to dismiss any submissions flagged by OpenAI's detection tool. Instead we decided to reject submissions only when clear evidence of copied and pasted text from external sources is identified. We do not reject essays in cases where keystrokes suggest low cognitive effort (e.g., continuous typing without revision), as this could occur not only when transcribing text from other sources but also when cognitive skills and working memory are sufficiently robust to generate responses on-the-fly while typing.

After completing the writing tasks, participants were asked to download their keystroke data from the site (stored locally for privacy reasons), and submit to us via Qualtrics. Subsequent approval and payment processing were administered on Prolific. We recruited 1,045 participants in three phases of crowdsourcing. The only difference in data collection was that after the first cohort we lowered the minimum essay length from 250 words to 150 words in order to attract more learners of English at lower proficiency levels.

Our motivation and focus for this research is on learners of English and their typing patterns at different levels of proficiency. Nevertheless we allowed native speakers of English to take part in the Prolific study, in order to have control data relating to how people type in English *in general*, so that we could identify patterns of typing which might be specific to learning English. Note that we view the native speakers in our study as controls not models for the learners. There is a long-running debate about putting native speakers on a pedestal in language teaching and assessment, with persuasive arguments against doing so (Phillipson, 1992; Cook, 1999; Alptekin, 2002) while others describe the potential value in presenting native speaker varieties to learners for context (Timmis, 2002; Adolphs, 2005).

W&I score	CEFR level
0	–
1	A1.i
2	A1.ii
3	A2.i
4	A2.ii
5	B1.i
6	B1.ii
7	B2.i
8	B2.ii
9	C1.i
10	C1.ii
11	C2.i
12	C2.ii
13	C2.ii

Table 1: How scores from the W&I automarker map to CEFR levels. Note that 13 is intended to map to C2.iii but we only used a scale of 0-12 and so map 13 to C2.ii.

Overall, 178 of our 1,045 participants self-identified as native speakers of English (approximately 17%); of the non-native speaker remainder, only 2% identified as beginners learning English, 38% identified as intermediate, and 60% identified as advanced.

5. Data Processing

As described above, participants declared themselves to be native speakers of English or not, and any non-native speakers of English were asked to report their proficiency level. Besides this information, we also obtained automatic grades for the submitted essays from the text API which is used by Write & Improve (W&I), an L2 learning tool that offers learners estimated grades and error feedback on their open writing. The W&I-specific generic multi-level CEFR grader/scorer for English texts estimates the language level on a scale from 0 to 13. Note that we round the floating point values from the automarker to the nearest whole number.

These integer scores can then be straightforwardly mapped to the CEFR scale, as shown in Table 1. Zero is a failure to register on the scale, 1 maps to A1.i, 2 maps to A1.ii, a score of 3 maps

to A2.i, and so on. In other words, there are two partitions within each of the six major CEFR levels. A top score of 13 was intended in the design of the automarker to mean C2.iii – for an essay even better than C2.ii. In our case we map a score of 13 to C2.ii because we did not ask human assessors to use this grade, since a bifurcation for each CEFR level is conceptually more straightforward than having bifurcations for levels A1-C1 then 3 grades for C2.

In parallel, we asked three qualified but non-operational assessors to grade the essays on the 0-12 scale described above (not using the maximum score of 13 used by the W&I automarker). It is important to note that this is not the usual way of arriving at a CEFR level for a learner of English: for one thing, language assessment is normally multi-skilled, including speaking, reading and listening skills, not just essay writing. For another, the essay prompts were 'just for fun', meaning that they are not in the usual style for a language exam and do not elicit the usual constructions and lexis. In addition, operational examiners most often assess texts submitted to an exam with a specific proficiency level and with assessment criteria relevant to that level. Applying a raw CEFR level to widely differing texts without reference to specific assessment criteria is unusual and challenging. Nevertheless, even with these caveats in mind, it gives us some information about text quality of the essays and how that might relate to keystroke patterns.

From carrying out this human annotation step, 35 of the 1,045 essays were rejected by at least one of the assessors for one of several reasons – the most common reasons were that the essay was off-topic for the given prompt, offensive or potentially distressing in some way. These 35 essays are not included in the public release of this dataset as it would be inappropriate to do so. Additionally, 4 submissions were subsequently removed after further review revealed that those participants were using a tablet, which was against our submission guidelines. Therefore the public release features 1,006 appropriate essays which were on topic and unaffected by the problems listed above. This is also the dataset we describe and analyse below.

6. Dataset Description

Following the acquisition of raw data from Qualtrics, we conducted a thorough data cleaning process to streamline the generation of three primary tables, conveniently saved in CSV format. The first table encompasses a wealth of submission details, including metadata and demographic information, along with the evaluations by three human assessors and the W&I automarker. Additionally, it features the original prompts, the final submitted text, and post-text processing attributes such as word counts per task and task completion times. A comprehensive explanation of the table's columns is available in the Appendix.

With regards to the keystroke log data, each JSON file associated with individual tasks was thoroughly processed and reformatted. This modification allowed for the generation of two comprehensive tables, showcasing the collective data of all participants in a convenient CSV format. Each row in the tables is precisely aligned with a distinct event, which is categorized as either user-related or system-related. This deliberate organization and representation of the data enhances its clarity and facilitates more accurate and efficient analysis. For each table, the fields available are:

- `id`: the user's anonymised number. This is to be used for the correspondence between the keystroke log data and the annotations.
- `time`: timestamp of the event, in milliseconds since the application was started.
- `type`: the event type. The possible values are `down` when a keyborard key is pressed, `up` when a keyboard key is released, `click` for a mouse click, `insert` when the content of the text-box is updated, and `capture`, which is periodically triggered to save the current text-box state.
- `key`: for key press and release events, this is the actual value of the key.
- `key_code`: for key press and release events, this is the name of the physical key on the keyboard rather than the specific layout chosen by the user. The values are associated with a US-based QWERTY layout.
- `alt_key`: indicator of whether an alt key is also pressed at the time of the event.
- `ctrl_key`: indicator of whether a control key is also pressed at the time of the event.
- `meta_key`: indicator of whether a meta key is also pressed at the time of the event.
- `shift_key`: indicator of whether a shift key is also pressed at the time of the event.
- `is_repeat`: indicator of whether this is a key that is automatically repeating because the key is held down.
- `range_start`: At the time of the event, the location as an integer offset in the text-box of either the cursor or the start of any selected text.
- `range_end`: At the time of the event, the location as an integer offset in the text-box of either the cursor or the end of any selected text.

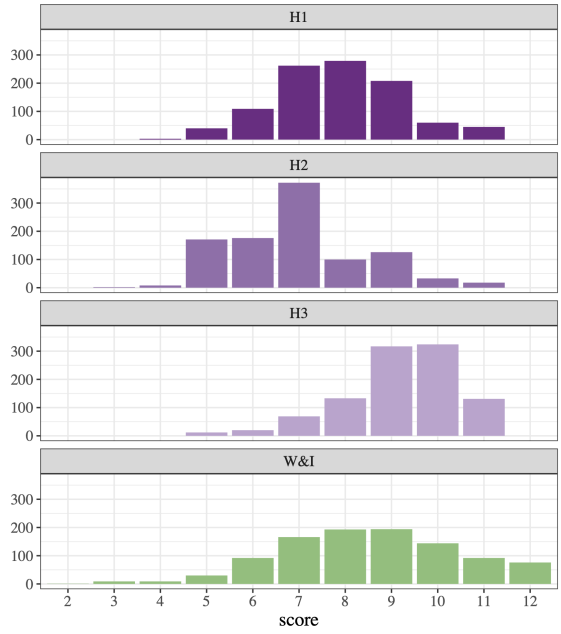


Figure 1: The distribution of marks for the 1,006 essays in the dataset, on a scale from 0 to 12 (scores of 0 and 1 do not feature), with subplots for each of the 3 human markers (H1, H2, H3) and the Write&Improve automarker (W&I).

- `text`: the text in the text-box when there is a `capture` event type. This may also be populated when there is an `insert` event type, with the content added to the text-box.

Note that the `key` values are useful for determining which keys are produced in the text editor. Values from `key_code` are useful to determine a user's keyboard layout and to sync keypress events with key release events; note that 'key' is not reliable for this task because the state of the alt/ctrl/meta/shift keys may have changed. Also, one cannot accurately determine the state of the alt/ctrl/meta/shift key modifiers from their Boolean values alone because a user may have caps-lock or alt-lock turned on. Their usage is primarily for use in understanding how a user is physically typing. A more detailed explanation is available in the Appendix.

7. Inter-annotator Agreement

To measure inter-annotator agreement on essay proficiency, we report root-mean-square deviation (RMSD), Spearman's rank correlation, and Gwet's agreement coefficient with quadratic weights (Gwet, 2002). This follows Yannakoudakis and Cummins (2015)'s recommendation for evaluating the performance of automated text scoring systems. Gwet's AC is a chance-adjusted agreement statistic similar to Cohen's kappa, except that it is based on expected disagreement rather than expected agreement. This mitigates a common problem with essay grading whereby imbalanced classes might lead to a low value of kappa even when agreement is high. Quadratic weighting is preferred so as to penalise larger disagreements more severely than those differing by only a single grade. The latter scenario can often result from essays near the borderline between grades.

	H1	H2	H3	W&I
H1	–	0.633	0.711	0.574
H2	0.633	–	0.567	0.563
H3	0.711	0.567	–	0.514
W&I	0.574	0.563	0.514	–
Avg	0.639	0.588	0.598	0.550

Table 2: Spearman's rank correlations between each marker: the three human assessors (H1, H2, H3) and the Write&Improve automarker (W&I). The average of correlations is also reported (Avg).

With the remaining 1,006 essays, Gwet's AC amongst the three markers and the automarker was 0.854. This is a good level of agreement; however we do find that the individual assessors have different marking traits. We show the distribution of marks by each of the human assessors and the automarker in Figure 1. It is apparent that one of the human assessors is relatively strict (H2), one is relatively lenient (H3) and the other is more evenly distributed around the middle of the marking scale (H1) similarly to the W&I automarker. This observation serves as a reminder as to why it is good practice to multi-mark for high stakes exams, and illustrates how automarkers can avoid some individual biases held by human markers.

Spearman's rank correlations for each pair of markers are shown in Table 2. The correlations are on the whole strong, in that they are all >0.5 and statistically significant ($p < 0.0001$), but it is noticeable that the judgements of the human assessors correlate with each other more than they do with the automarker. This could be a reason to ignore the automarker scores, except that we also find H2 and H3 correlating with each other to the same degree as the human markers do with the automarker. It is clear that H1 correlates most strongly with the other markers and the W&I automarker.

Finally, we calculate RMSD between each marking pair, including the three human markers and W&I automarker. We also show the RMSD compared to H_{avg} – the average of human marks (i.e. not including the automarker). We use RMSD rather than root-mean-square *error* because there is no clear 'ground truth' in terms of essay assessments, but rather we are working towards it. There-

	H_{avg}	H1	H2	H3
H1	0.622	-	-	-
H2	1.288	1.487	-	-
H3	1.371	1.695	2.586	-
W&I	1.543	1.708	2.241	1.770

Table 3: Pairwise RMSD between all markers. H_{avg} denotes the average mark received from all human assessors.

A2	B1	B2	C1	C2
4	111	549	323	19
0.4%	11%	54.6%	32.1%	1.9%

Table 4: The count of approved essays in the dataset by CEFR level (row 1), calculated as the mean scores from three human assessors and the Write&Improve automarker. In row 2 we show the proportional distribution of essays by CEFR level out of the total 1,006 essays in the dataset. There were no essays deemed to be level A1.

fore we opt for RMSD in order to report on the level of *divergence* in judgements. Pairwise RMSDs are shown in Table 3 where values of less than 1 indicate that on average the pair deviate by less than a micro-CEFR level in their assessments. Values greater than 1 indicate that the pair tend to deviate by more than a micro-CEFR level, and values greater than 2 represent a whole CEFR level's difference in judgements across the dataset. We find that H1 is closest to the mean of human marks and has the lowest deviation from other markers including the automarker. H3 diverges most from the human average but it is H2 who is involved in the highest RMSD value of all (with H3) and has the highest RMSD with the W&I automarker.

Finally, we calculate the mean of the four scores for each essay – from the three human assessors and the automarker – and include these in the dataset release rounded to the nearest integer from 0 to 12, and mapped to the appropriate CEFR grade. We show the distribution of the six macro CEFR levels in the essays dataset in Table 4. It is evident that the majority were assessed to be upper intermediate level (CEFR B2) with the next biggest tranche being advanced (C1). A small number were lower intermediate (B1) and 'proficiency' (C2). Only a few were 'basic level' A2 and none were beginner A1.

8. Data Analysis

Regarding demographic data, Polish was the most commonly reported native language among our participants, constituting 20% of the responses. It was closely followed by English and Portuguese, each comprising 17% of the total. Other represented native languages included Spanish (12%), Italian (9%), and Greek (5%), with the remaining participants reporting a variety of 37 other languages. In terms of age distribution, 25% of the participants were between 18 and 23 years old, and the median age was 26 years. Furthermore, 25% of the participants were above 32 years old. In relation to keyboard usage, a significant majority of 89% of the participants reported using a QWERTY keyboard layout, while the QWERTZ layout was the next most popular, used by 7% of respondents.

Additionally, we compared the survey completion duration between native English speakers (NS) and non-native English speakers (NNS) using the Kolmogorov-Smirnov (K-S) test. The calculated K-S statistic was found to be 0.10, with a p-value of 0.10, suggesting no significant difference between NS and NNS in terms of the time required to complete the survey. Similar results were observed when comparing the task-specific completion times, indicating that NS did not generally complete the tasks more quickly than NNS (KS $= 0.1, p = 0.08$). Contrarily, a significant difference was observed for the average marks received (KS $= 0.47, p < 0.001$), where NS achieved higher marks, as expected. To assess the relevance of essay scores to the prompt type, pairwise K-S tests were performed on the marks received across different prompts. These tests revealed that the prompt *Write a short story with the title "Unforgettable"* resulted in higher essay scores amongst other prompts.

Moreover, we observed a weak correlation between the time spent on the essay task and the average mark received, both from human markers and the automarker ($r = 0.09, p = 0.003$). In contrast, we observed a moderate negative correlation between the time they spent on the copy-text task and the average mark received on the essay-writing task ($r = -0.27, p < 0.001$), implying that fast typists could generally achieve higher marks. Additionally, a moderate correlation was noted between the average mark received on the essay task and the number of keystrokes ($r = 0.43, p < 0.001$), and a stronger correlation was evident between the number of words and the average CEFR score ($r = 0.51, p < 0.001$), aligning with expectations based on previous studies (Crossley et al., 2011; McNamara et al., 2015; Ke and Ng, 2019; Ramesh and Sanampudi, 2022). Similar correlation results are shown in Table 5, while Figure 2 provides a visual representation of some aspects of our analysis.

Finally, we conducted statistical testing at the keystroke level. More specifically, for each user and for each task, we computed the key press latency t_{PL} (i.e., the time interval between consecu-

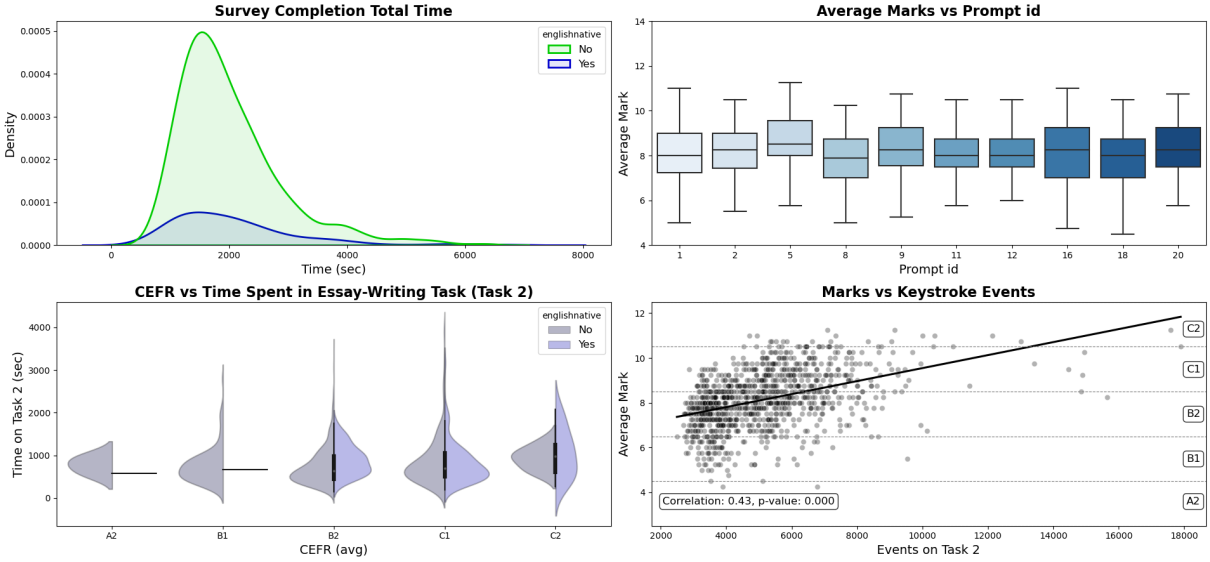


Figure 2: **Top left**: density of total completion time for native English speakers and non-native English Speakers. **Top right**: Average marks for different prompt ids. **Bottom left**: Estimations of the distributions of the time spent on task 2 for distinct CEFR scores achieved. **Bottom right**: Pearson's correlation between the number of events on task 2 and the average mark.

Variable	r
Age	+0.10**
Hours spent on a computer per day	-0.04
Hours exposed to English per day	+0.18***
Number of years learning English	+0.18***
Time needed to complete the survey	-0.06
Time needed to complete task 1	-0.27***
Time needed to complete task 2	+0.09**
Number of keystroke log events in task 2	+0.43***
Number of words in task 2	+0.51***

Table 5: Pearson's correlation coefficients between different variables and the average mark received by the human markers and the automarker. 2 asterisks indicate p values less than 0.01; 3 asterisks indicate p values less than 0.001.

tive `key down` events). For each participant, the distribution of t_{PL} samples for the copy-text task is expected to be different from that of the essay-writing task, due to the anticipated greater cognitive load in the latter. In fact, one would expect the distribution of the essay-writing task to be bimodal due to pauses while thinking, and bursts while typing (Locklear et al., 2014; Baaijen and Galbraith, 2018; Conijn et al., 2021; González et al., 2021). Since the tails of these distributions are important, we decided to perform an Anderson-Darling test. Participants for which the corresponding p-values are high, indicate cases where the copy-text task and the essay-writing task are not statistically different in terms of the key press latency distribution, thus they could subsequently indicate cases where the answers are either memorised or transcribed from external sources. As described earlier, we decide to still keep those submissions in our dataset but highlight the importance of further considerations. Notably, 200 submissions exhibited a p-value exceeding 0.05. Additional details can be found in the Appendix.

9. Conclusions & Future Work

In this paper we have introduced the new KUPA-KEYS dataset which includes keystroke data from 1,006 participants. The participants were native speakers and non-native speakers of English who wrote two texts: a copy-text task of 300 words, and an essay-writing task responding to one of ten prompts. The dataset also includes metadata about the individual participants such as age, location, level of English and other languages known. In addition our JavaScript keylogger is open-source, and we make our data collection website freely available for the sake of reproducibility.

We annotated the essays with proficiency assessments from both human assessors and a pre-trained automarker. We found a decent level of agreement amongst these assessors, while also finding different tendencies, and use the average of the scores to determine an approximate proficiency level for each essay on the CEFR scale. Our initial analyses revealed that the time spent on the task only weakly correlated with the mark received, whilst the number of keystrokes (and thereby number of words) held a stronger corre-

lation. The dataset carries the potential for further analyses of typing patterns, indications of complex word and character sequences, and identification of hierarchical structures in the writing process per Ballier et al. (2019) and Leijten et al. (2019).

In parallel we are working on visualisation techniques to show pauses and chunks in student writing, as teaching aids for language learners. We note that pauses can indicate a number of author behaviours, including reflection, distraction, and difficulty in finding the right word or phrase to continue composition (Banerjee et al., 2014). If we can distinguish between these types of pauses, then it may enable writing assistance through well-timed interventions. For instance, if we can successfully identify when writers are struggling with linguistic constructions then we can offer writing assistance accordingly (Conijn et al., 2021). This support could be in the form of writing suggestions, a chatbot or dictionary look up tools. In some cases we can make this kind of writing support pedagogically useful for learners of English.

Other future work includes the use of features derived from keystroke data to enhance essay assessment models (Chukharev-Hudilainen, 2019). Features have been used in 'classical' machine learning approaches, such as linear regression, decision trees, and so on, but to the best of our knowledge keystroke information has not been incorporated into Transformer-based assessment models of the kind currently being developed (Mizumoto and Eguchi, 2023). One reason we wish to release this dataset is to enable others to work with it on projects such as the ones described in this section.

Finally, the continuing challenge of generative-AI text detection is acknowledged in recent literature (Krishna et al., 2024; Sadasivan et al., 2023), signifying a necessity for increased endeavour in this domain. Moreover, it is plausible to hypothesize that, given our dataset comprises the complete history of keystroke log data, one might explore the absence of cognitive effort exhibited during text production – a phenomenon evident when text is transcribed from alternative sources. For instance when comparing the typing speeds for the two tasks – one a copy task and the other creative writing – the difference is stark: a mean in our dataset of 37 words per minute for the former and 21 words per minute for the latter. We are currently reviewing the keystroke data to identify suspicious bursts of text or copy-paste events which might be signs of plagiarism from other sources including generative AI. We will release the scripts and annotation from this review in a later update to the public dataset.This exploration may potentially facilitate the advent of novel research towards generative-AI text detection based on event-based information including keystrokes.

10. Ethics & Limitations

It is important to consider privacy concerns when collecting keylogging data. In addition to academic applications, keylogger applications also feature prominently in many illegal data hacking attempts and questionable marketing tactics by large corporations. Our JavaScript data capture interface works by storing all of a typist's data locally on a user's machine in what is called the Document Object Model (DOM). No data is automatically sent to a remote server. Thus the user could download their keystroke data locally and had to agree to submit it to us. Future work in this area may involve remote keystroke logging which brings ethical implications, including the need to properly inform participants as to the data being collected, and the need to safely store it. One possible ethical solution for academic research and educational applications could be to immediately extract the kind of keystroke metrics and features discussed in this paper and discard the raw data.

We are interested in future work exploring cognitive models, linguistic analyses (Pacquetet, 2024) and educational applications for keystroke data. We believe that the first two are justified as scientifically interesting areas of research; with regard to educational applications, whenever collecting data from end-users it is important to do so in an ethical way that allows individuals to opt out if they wish, and to make use of the data in ways that we can demonstrate are useful in the short-term, non-intrusive, and potentially enhance learning in the long-term.

Acknowledgements

We thank Ece Washbrook & Russell Moore for their previous work on keystroke logging in an internship project at the University of Cambridge; similarly we thank Souradj Mounien Dit Ravi for his internship project at Université Paris Cité. This work was supported by a research grant from Université Paris Cité and King's College London as part of the Deep Learning for Language Assessment (DLLA) project, under the ANR grant ANR-18-IDEX-0001, Financement IdEx Université de Paris. In addition the ALTA Institute is supported by Cambridge University Press & Assessment. Finally we thank ELiT for use of the prompts, the human annotation of essay scores, and access to the W&I API, Øistein Andersen for discussion around the W&I scoring scale and mapping that to CEFR levels, and Mark Elliott for discussion around IAA.

Bibliographical References

Alejandro Acien, Aythami Morales, Ruben Vera-Rodriguez, Julian Fierrez, and John V. Monaco. 2020. TypeNet: Scaling up keystroke biometrics. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–7.

Svenja Adolphs. 2005. "I don't think I should learn all this" – a longitudinal view of attitudes towards 'native speaker' English. In Claus Gnutzmann and Frauke Intemann, editors, *The globalisation of English and the English language classroom*, pages 119–131. Tübingen: Narr Verlag.

Jeffrey D Allen. 2010. *An analysis of pressure-based keystroke dynamics algorithms*. Ph.D. thesis, Southern Methodist University.

Cem Alptekin. 2002. Towards intercultural communicative competence in ELT. *ELT Journal*, 56:57–64.

Veerle M Baaijen and David Galbraith. 2018. Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction*, 36(3):199–223.

Yukino Baba and Hisami Suzuki. 2012. How are spelling errors generated and corrected? a study of corrected and uncorrected spelling errors using keystroke logs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 373–377, Jeju Island, Korea. Association for Computational Linguistics.

Nicolas Ballier, Erin Pacquetet, and Taylor Arnold. 2019. Investigating Keylogs as Time-Stamped Graphemics. In *Proceedings of Graphemics in the 21st Century, Brest 2018*, pages 353–365.

Ritwik Banerjee, Song Feng, Jun Seok Kang, and Yejin Choi. 2014. Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1469–1473.

Evgeny Chukharev-Hudilainen. 2019. Empowering automated writing evaluation with keystroke logging. In *Observing writing*, pages 125–142. Brill.

Evgeny Chukharev-Hudilainen, Aysel Saricaoglu, Mark Torrance, and Hui-Hsien Feng. 2019. Combined deployable keystroke logging and eyetracking for investigating L2 writing fluency. *Studies in Second Language Acquisition*, 41(3):583–604.

Rianne Conijn, Emily Dux Speltz, and Evgeny Chukharev-Hudilainen. 2021. Automated extraction of revision events from keystroke data. *Reading and Writing*, pages 1–26.

Vivian Cook. 1999. Going beyond the native speaker in language teaching. *TESOL Quarterly*, 33:185–209.

Scott A. Crossley, Jennifer L. Weston, Susan T. McLain Sullivan, and Danielle S. McNamara. 2011. The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, 28(3):282–311.

Vivek Dhakal, Anna Maria 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*, pages 1–12.

Anna Maria Feit, Daryl Weir, and Antti Oulasvirta. 2016. How we type: Movement strategies and performance in everyday typing. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4262–4273.

Julian Fierrez, Javier Galbally, Javier Ortega-Garcia, Manuel R Freire, Fernando Alonso-Fernandez, Daniel Ramos, Doroteo Torre Toledano, Joaquin Gonzalez-Rodriguez, Juan A Siguenza, Javier Garrido-Salas, et al. 2010. Biosecurid: a multimodal biometric database. *Pattern Analysis and Applications*, 13(2):235–246.

David Galbraith and Veerle Baaijen. 2019. Aligning keystrokes with cognitive processes in writing. In E. Lindgren and K. Sullivan, editors, *Observing writing: Insights from keystroke logging and handwriting*, pages 306–325. Leiden: Brill.

Romain Giot, Mohamad El-Abed, and Christophe Rosenberger. 2009. GREYC keystroke: a benchmark for keystroke dynamics biometric systems. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6.

Romain Giot, Mohamad El-Abed, and Christophe Rosenberger. 2012. Web-based benchmark for keystroke dynamics biometric systems: A statistical analysis. In *2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 11–15. IEEE.

Nahuel González, Enrique P. Calot, Jorge S. Ierache, and Waldo Hasperué. 2021. On the shape

of timings distributions in free-text keystroke dynamics profiles. *Heliyon*, 7(11):e08413.

Adam Goodkind and Andrew Rosenberg. 2015. Muddying the multiword expression waters: How cognitive demand affects multiword expression production. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 87–95, Denver, Colorado. Association for Computational Linguistics.

Daniele Gunetti and Claudia Picardi. 2005. Keystroke analysis of free text. *ACM Transactions on Information and System Security (TISSEC)*, 8(3):312–347.

Kilem Gwet. 2002. Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*, 2:1–9.

Syed Zulkarnain Syed Idrus, Estelle Cherrier, Christophe Rosenberger, and Patrick Bours. 2013. Soft biometrics database: A benchmark for keystroke dynamics biometric systems. In *2013 International Conference of the BIOSIG Special Interest Group (BIOSIG)*, pages 1–8. IEEE.

Pilsung Kang and Sungzoon Cho. 2015. Keystroke dynamics-based user authentication using long and free text strings from various input devices. *Information Sciences*, 308:72–93.

M. Karnan, M. Akila, and N. Krishnaraj. 2011. Biometric personal authentication using keystroke dynamics: A review. *Applied Soft Computing*, 11(2):1565–1573. The Impact of Soft Computing for the Progress of Artificial Intelligence.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.

John Leggett, Glen Williams, Mark Usnick, and Mike Longnecker. 1991. Dynamic identity verification via keystroke characteristics. *International Journal of Man-Machine Studies*, 35(6):859–870.

Mariëlle Leijten, Eric Van Horenbeeck, and Luuk Van Waes. 2019. Analysing keystroke logging data from a linguistic perspective. In E. Lindgren and K. Sullivan, editors, *Observing writing: Insights from keystroke logging and handwriting*, page 71–95. Leiden: Brill.

Hilbert Locklear, Sathya Govindarajan, Zdeňka Sitová, Adam Goodkind, David Guy Brizan, Andrew Rosenberg, Vir V. Phoha, Paolo Gasti, and Kiran S. Balagani. 2014. Continuous authentication with cognition-centric text production and revision features. In *IEEE International Joint Conference on Biometrics*.

Cerstin Mahlow. 2015. Learning from Errors: Systematic Analysis of Complex Writing Errors for Improving Writing Technology. In *Language Production, Cognition, and the Lexicon*, pages 419–438.

Cerstin Mahlow, Malgorzata Anna Ulasik, and Don Tuggener. 2022. Extraction of transforming sequences and sentence histories from writing process data: a first step towards linguistic modeling of writing. *Reading and Writing*, pages 1–40.

Emanuele Maiorana, Himanka Kalita, and Patrizio Campisi. 2021. Mobile keystroke dynamics for biometric recognition: An overview. *IET Biometrics*, 10(1):1–23.

Danielle S. McNamara, Scott A. Crossley, Rod D. Roscoe, Laura K. Allen, and Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

John V Monaco, Ned Bakelman, Sung-Hyuk Cha, and Charles C Tappert. 2012. Developing a keystroke biometric system for continual authentication of computer users. In *2012 European Intelligence and Security Informatics Conference*, pages 210–216.

John V Monaco, Gonzalo Perez, Charles C Tappert, Patrick Bours, Soumik Mondal, Sudalai Rajkumar, Aythami Morales, Julian Fierrez, and Javier Ortega-Garcia. 2015. One-handed keystroke biometric identification competition. In *2015 International Conference on Biometrics (ICB)*, pages 58–64. IEEE.

John V Monaco, John C Stewart, Sung-Hyuk Cha, and Charles C Tappert. 2013. Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8.

John V. Monaco and Charles C. Tappert. 2016. Obfuscating keystroke time intervals to avoid

identification and impersonation. *CoRR*, abs/1609.07612.

Jugurta R Montalvão Filho and Eduardo O Freire. 2006. On the equalization of keystroke timing histograms. *Pattern Recognition Letters*, 27(13):1440–1446.

Mozilla Foundation. 2023. Mozilla WebDocs: High precision timing. https://developer.mozilla.org/en-US/docs/Web/API/Performance_API/High_precision_timing. Accessed: 2023-10-17.

Christopher Murphy, Jiaju Huang, Daqing Hou, and Stephanie Schuckers. 2017. Shared dataset on natural human-computer interaction to support continuous authentication research. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 525–530. IEEE.

Erin Pacquetet. 2024. *The effect of linguistic properties on typing behaviors and production processes*. Ph.D. thesis, University at Buffalo.

Alen Peacock, Xian Ke, and Matthew Wilkerson. 2004. Typing patterns: A key to user identification. *IEEE Security and Privacy*, 2(5):40–47.

Robert Phillipson. 1992. *Linguistic Imperialism*. Oxford University Press, Oxford, UK.

Barbara Plank. 2016. Keystroke dynamics as signal for shallow syntactic parsing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 609–619.

Barbara Plank. 2018. Predicting authorship and author traits from keystroke dynamics. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 98–104, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Giorgio Roffo, Cinzia Giorgetta, Roberta Ferrario, Walter Riviera, and Marco Cristani. 2014. Statistical analysis of personality and identity in chats using a keylogging platform. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 224–231.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-generated text be reliably detected? arXiv:2303.11156.

Moritz Jonas Schaeffer, Michael Carl, Isabel Lacruz, and Akiko Aizawa. 2016. Measuring cognitive translation effort with activity units. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 331–345.

John C Stewart, John V Monaco, Sung-Hyuk Cha, and Charles C Tappert. 2011. An investigation of keystroke and stylometry traits for authenticating online test takers. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–7.

Giuseppe Stragapede, Paula Delgado-Santos, Ruben Tolosana, Ruben Vera-Rodriguez, Richard Guest, and Aythami Morales. 2023. Mobile keystroke biometrics using transformers. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6.

Yan Sun, Hayreddin Ceker, and Shambhu Upadhyaya. 2016. Shared keystroke dataset for continuous authentication. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

Charles C Tappert, Mary Villani, and Sung-Hyuk Cha. 2010. Keystroke biometric identification and authentication on long-text input. In *Behavioral biometrics for human identification: Intelligent applications*, pages 342–367.

Pin Shen Teh, Andrew Beng Jin Teoh, Connie Tee, and Thian Song Ong. 2011. A multiple layer fusion approach on keystroke dynamics. *Pattern Analysis and Applications*, 14(1):23–36.

Ivor Timmis. 2002. Native-speaker norms and International English: a classroom view. *ELT Journal*, 56:240–249.

Vishaal Udandarao, Mohit Agrawal, Rajesh Kumar, and Rajiv Ratn Shah. 2020. On the inference of soft biometrics from typing patterns collected in a multi-device environment. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 76–85.

Esra Vural, Jiaju Huang, Daqing Hou, and Stephanie Schuckers. 2014. Shared research dataset to support development of keystroke authentication. In *IEEE International joint conference on biometrics*, pages 1–8. IEEE.

Helen Yannakoudakis and Ronan Cummins. 2015. Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223.

Mo Zhang, Jiangang Hao, Chen Li, and Paul Deane. 2016. Classification of writing patterns using keystroke logs. In *Quantitative psychology research*, pages 299–314. Springer.

A. Appendix

A.1. Data Collection

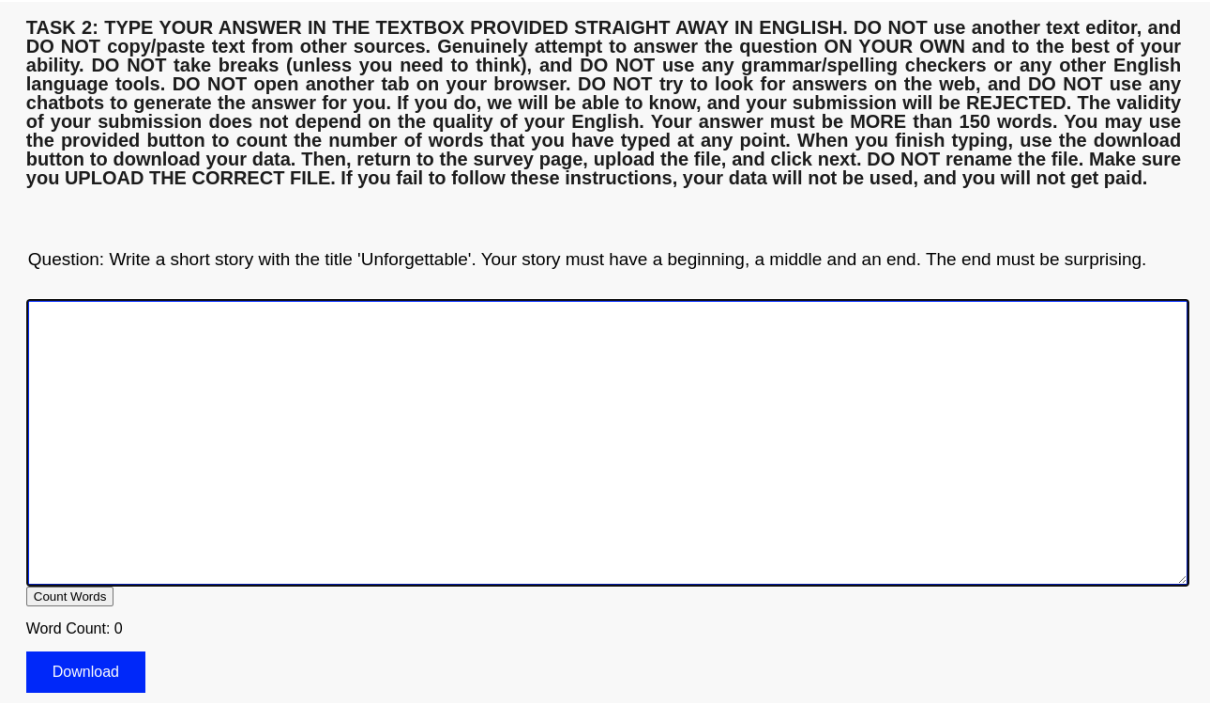


Figure 3: Screenshot of the data collection interface for task 2, the essay writing task.

A.2. Dataset Description

The table `KUPA-KEYS-META.csv` of demographic data, metadata, and pre-processed data extracted from the keylogger includes the following 38 columns:

id: The unique user id that participated in the survey. This is not connected with the participants' Prolific identification number. It is the submission number extracted from Qualtrics.

navigator_useragent: This value is extracted from the keylogger application and contains information about the user's web browser and device.

navigator_language: This value extracted from the keylogger application and contains information about the user's preferred language setting in their web browser. This setting is typically determined based on the user's browser preferences or system settings.

age: Response to the question "What is your age in years?". The value is an integer in [18,99].

handedness: This is the participants' response to the question "Are you right-handed, left-handed, or ambidextrous". The value is in the set Right-handed, Left-handed, Ambidextrous.

comphours: Response to the question "How many hours per day do you spend on a computer?". The value is an integer in [0,24].

layoutcomf: Response to the question "Which keyboard layout are you most comfortable with?". The value can be QWERTY, QWERTZ, AZERTY, IDK (I don't know), or Other (specified in a text box).

layoutnow: Response to the question "Which keyboard layout are you using now?". As previously, the value can be QWERTY, QWERTZ, AZERTY, IDK (I don't know), or Other (specified in a text box).

comptype: Response to the question ”Are you using a desktop or a laptop now?”. The value is in Desktop, Laptop. Note that in this survey we only allowed using a desktop computer or a laptop.

countryres: Response to the question ”What is your current country of residence?”. A full comprehensive list of countries was provided as a drop down list.

nativelang: Response to the question ”What is your native language?”. A full comprehensive list of countries was provided as a drop down list.

englishyears: Response to the question ”How many years have you been learning English?”. This question was provided to non-native English speakers only. The value is a float number.

englishcountrymonths: Response to the question ”How much time (in months) have you spent in an English-speaking country over the last 3 years?”. This question was provided to non-native English speakers only. The value is an integer in [0,36].

englishexposure: Response to the question ”How many hours per day are you exposed to English on average?”. This question was provided to non-native English speakers only. The value is an integer in [0,24].

otherlanguages: Response to the question ”Besides your native language and English, what other languages do you speak? Choose the language you are most comfortable with”. A full comprehensive list of countries was provided as a drop down list.

cefrself: Response to the question ”Do you consider yourself a beginner (A1, A2), intermediate (B1, B2), or advanced (C1, C2) user of English?”. This question was provided to non-native English speakers only. The value is in the set 'Advanced (C1, C2)', 'Beginner (A1, A2)', 'Intermediate (B1, B2)', or null in the case of native English speakers.

cefrlevel: Response to the question ”If you have passed an English language exam, please let us know your highest level on the CEFR scale”. The value is in the set 'I have not taken an English language exam', 'My CEFR level is A1 or A2', 'My CEFR level is B1 or B2', 'My CEFR level is C1 or C2'.

cefrwhen: Response to the question ”In which year did that assessment take place?”. This question was provided only to participants who reported they've passed an English language exam in the previous question.

educ: Response to the question ”What is the highest degree or level of education you have completed?”. The options provided were ”High School”, ”Bachelor's Degree (e.g., BSc, BA, MB, etc.)” ”Master's Degree (e.g., MSc, MEng, MRes, etc.)”, ”Doctoral Degree or higher (e.g., PhD, MPhil, etc.)”, or ”Other (please specify)”.

time: The total time (in seconds) the participants needed to complete the survey, including the two tasks.

task1_time: The time (in seconds) the participants spent in the copy-text task.

task2_time: The time (in seconds) the participants spent in the essay-writing task.

task1_events: The number of events recorded by the keylogger for the copy-text task. Note that this figure includes both user-related events and system-related events.

task2_events: The number of events recorded by the keylogger for the essay-writing task. Note that this figure includes both user-related events and system-related events.

task1_words: The number of words in the final submitted text for the copy-text task.

task2_words: The number of words in the final submitted text for the essay-writing task.

cohort: The cohort number the participant. There were three cohorts. For the first cohort, the minimum number of words for the essay-writing task was specified as 250. This cohort also includes native English speakers. For cohorts 2 and 3, the minimum number of words was reduced to 150. There is no other difference between these cohorts, other than the season of when the survey took place.

prompt_id: An integer specifying the prompt id for the essay-writing task. Each participant was randomly allocated to a prompt out of ten options.

prompt: The actual essay prompt (as text).

answer: The final submitted answer of the user in the text box.

mark_a0: The mark of the automarker (W&I) as an integer in [0,12] on the submitted essay.

mark_h1: The mark of the first human marker as an integer in [0,12] on the submitted essay.

mark_h2: The mark of the second human marker as an integer in [0,12] on the submitted essay.

mark_h3: The mark of the third human marker as an integer in [0,12] on the submitted essay.

cefr_a0: The CEFR level predicted by the automarker on the submitted essay.

cefr_h1: The CEFR level from the first human marker on the submitted essay.

cefr_h2: The CEFR level from the second human marker on the submitted essay.

cefr_h3: The CEFR level from the third human marker on the submitted essay.

For the copy-text task and the essay-writing task we provide two more tables, `KUPA-KEYS-TASK-1.csv` and `KUPA-KEYS-TASK-2.csv`. For each table, each row corresponds to an event and includes the 13 following columns:

id: The participant's identification anonymised number. This is to be used for the correspondence between the keystroke log data and the entries in the table `KUPA-KEYS-META.csv`.

time: timestamp of the event, in milliseconds since the application was started.

type: the event type. The possible values are 'down' when a keyborard key is pressed, 'up' when a keyboard key is released, 'click' for a mouse click, 'insert' when the content of the text-box is updated, and 'capture', which is periodically triggered to save the current textbox state.

key: for key press and release events, this is the actual value of the key pressed.

key_code: for key press and release events, this is the name of the physical key on the keyboard rather than the specific layout chosen by the user. The values are associated with a US-based QWERTY layout.

alt_key: indicator of whether an alt key is also pressed at the time of the event.

ctrl_key: indicator of whether a control key is also pressed at the time of the event.

meta_key: indicator of whether a meta key is also pressed at the time of the event.

shift_key: indicator of whether a shift key is also pressed at the time of the event.

is_repeat: indicator of whether this is a key that is automatically repeating because the key is held down.

id	time	type	key	key_code	alt_key	ctrl_key	meta_key	shift_key	is_repeat	range_start	range_end	text
xa2	563.4	down	'I'	KeyI	-	-	-	True	-	0	0	-
xa2	564.7	capture	-	-	-	-	-	-	-	-	-	'I'
xa2	564.7	input	-	-	-	-	-	-	-	1	1	'I'
xa2	691.6	up	'I'	KeyI	-	-	-	True	-	1	1	-
xa2	708.0	up	'Shift'	ShiftLeft	-	-	-	-	-	1	1	-
xa2	708.3	down	' '	Space	-	-	-	-	-	1	1	-
xa2	709.6	input	-	-	-	-	-	-	-	2	2	' '
xa2	835.6	up	' '	Space	-	-	-	-	-	2	2	-

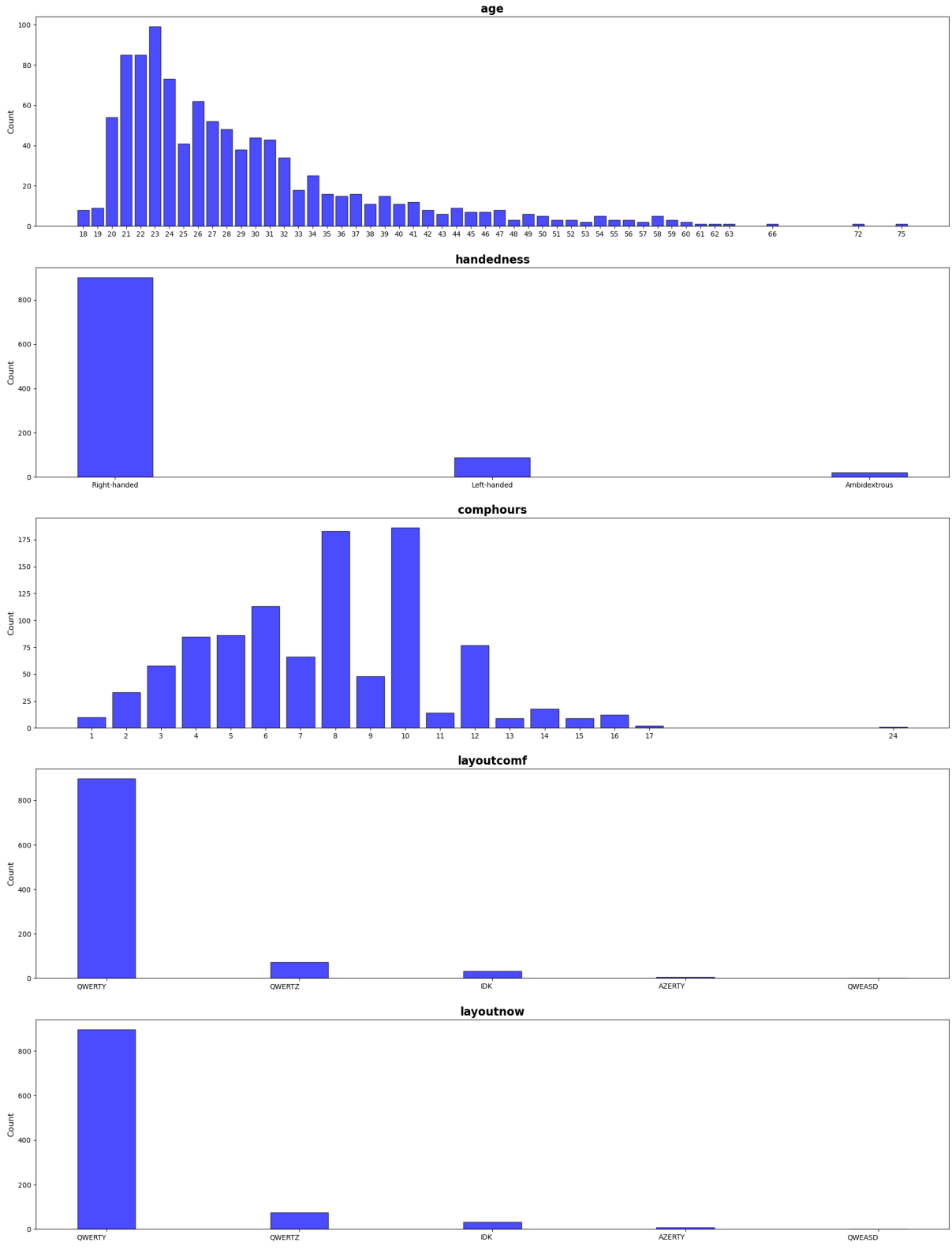
Table 6: Part of the keystroke log data for user with id xa2

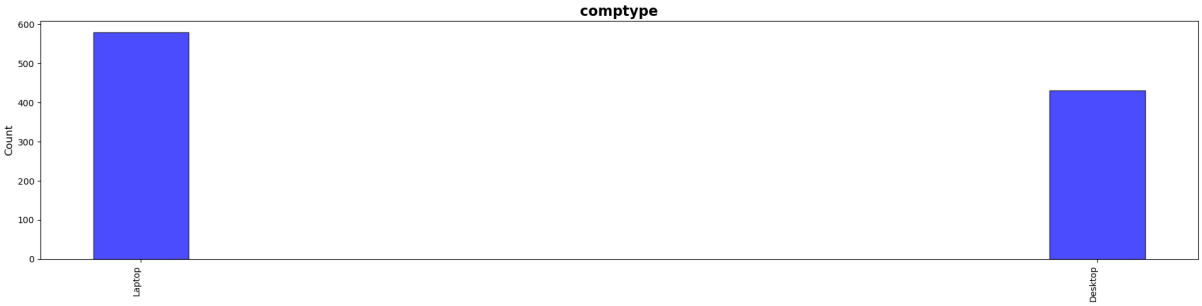
range_start: At the time of the event, the location as an integer offset in the text-box of either the cursor or the start of any selected text.

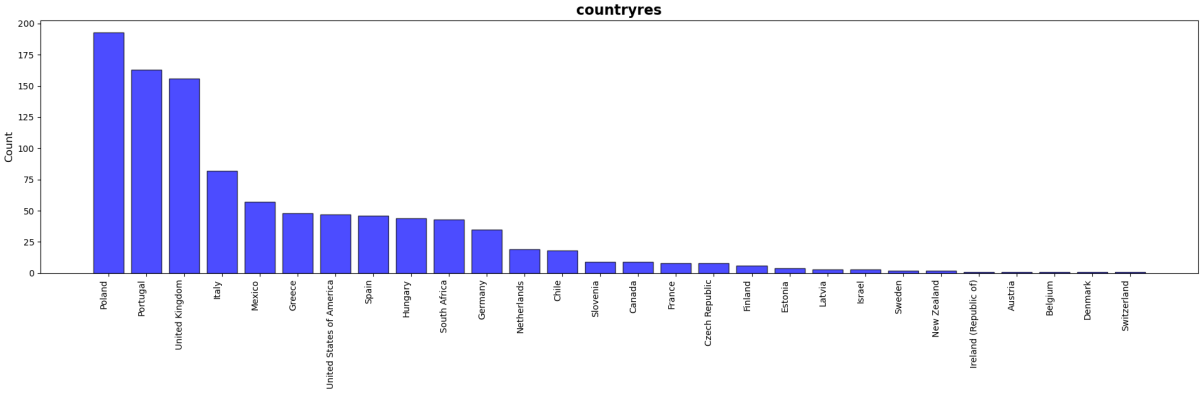
range_end: At the time of the event, the location as an integer offset in the text box of either the cursor or the end of any selected text.

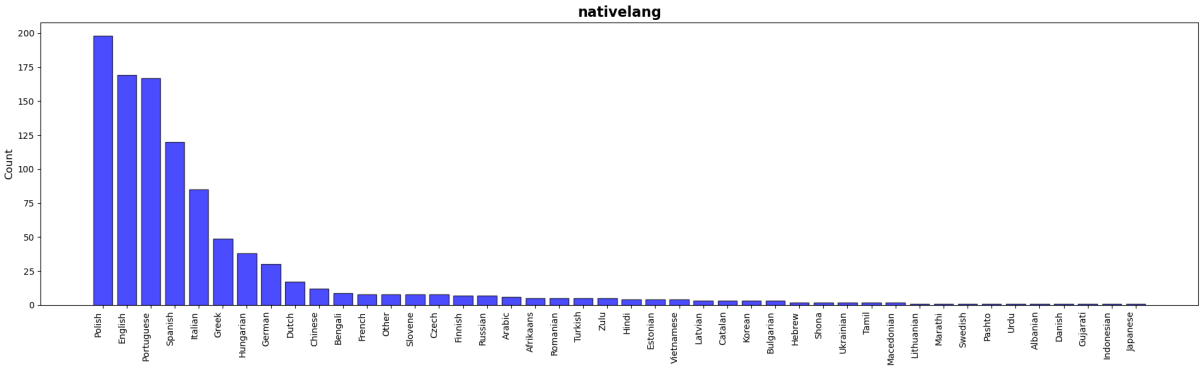
text: the text in the text box when there is a capture event type. This may also be populated when there is an insert event type, with the content added to the text box.

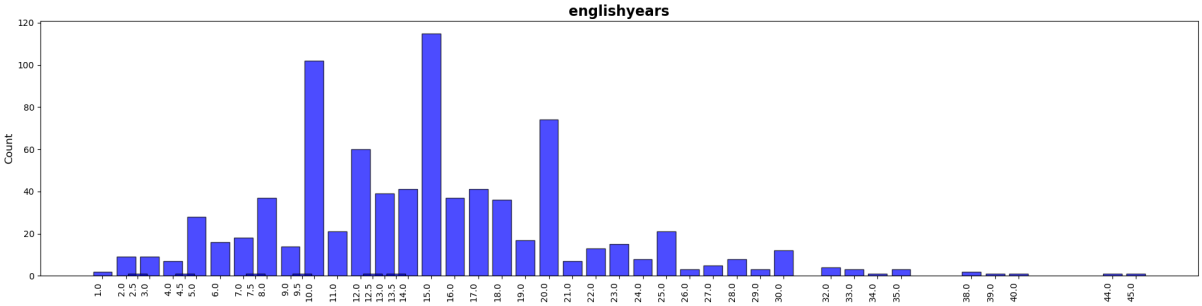
A.3. Demographic Data

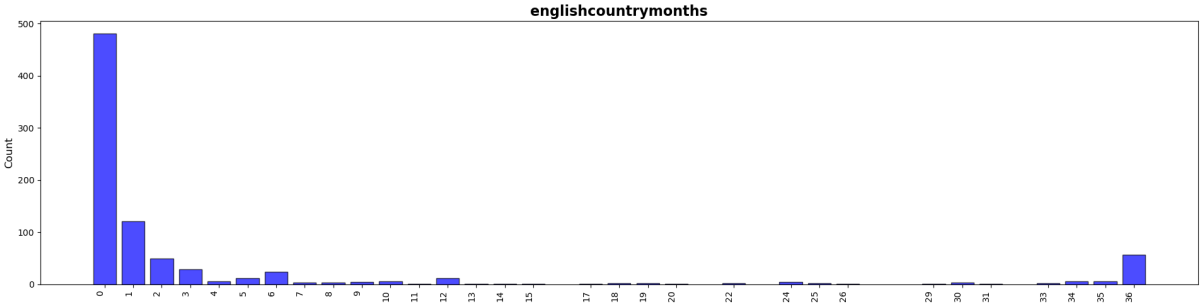


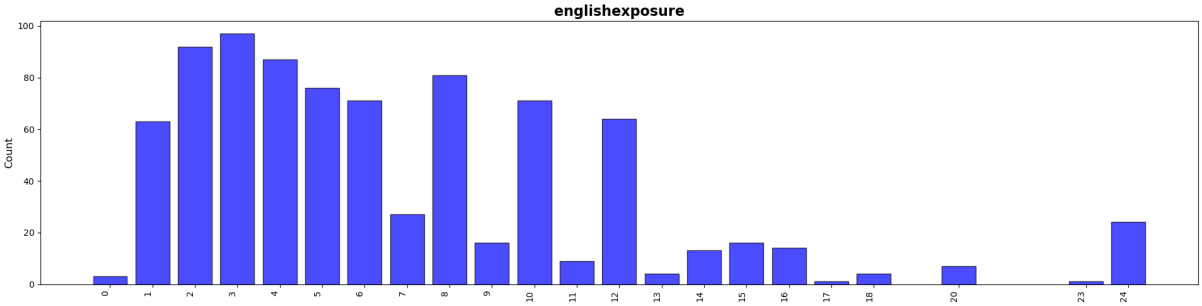

comptype
Count
Laptop
Desktop

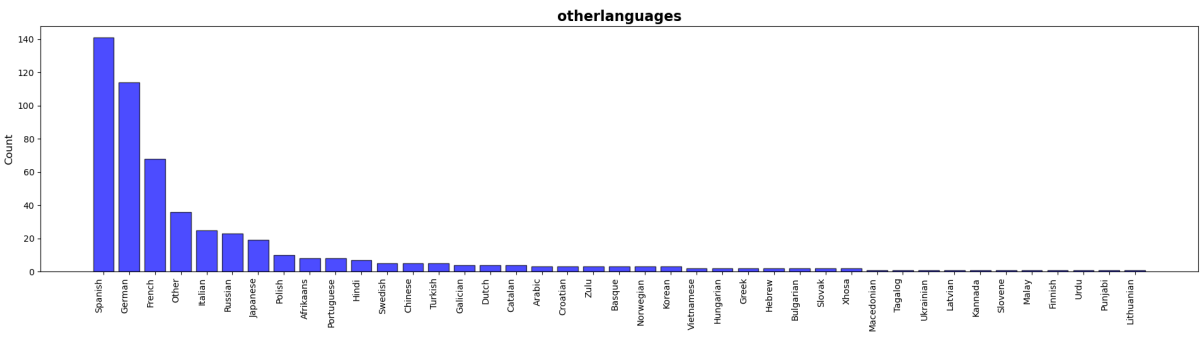

countryres
Count
Poland
Portugal
United Kingdom
Italy
Mexico
Greece
United States of America
Spain
Hungary
South Africa
Germany
Netherlands
Chile
Slovenia
Canada
France
Czech Republic
Finland
Estonia
Latvia
Israel
Sweden
New Zealand
Ireland (Republic of)
Austria
Belgium
Denmark
Switzerland

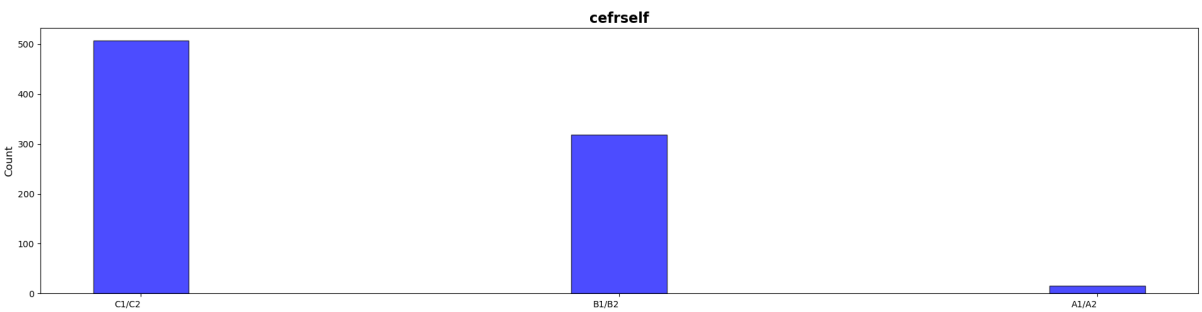

nativelang
Count
Polish
English
Portuguese
Spanish
Italian
Greek
Hungarian
German
Dutch
Chinese
Bengali
French
Other
Slovene
Czech
Finnish
Russian
Arabic
Afrikaans
Romanian
Turkish
Zulu
Hindi
Estonian
Vietnamese
Latvian
Catalan
Korean
Bulgarian
Hebrew
Shona
Ukrainian
Tamil
Macedonian
Lithuanian
Marathi
Swedish
Pashto
Urdu
Albanian
Danish
Gujarati
Indonesian
Japanese

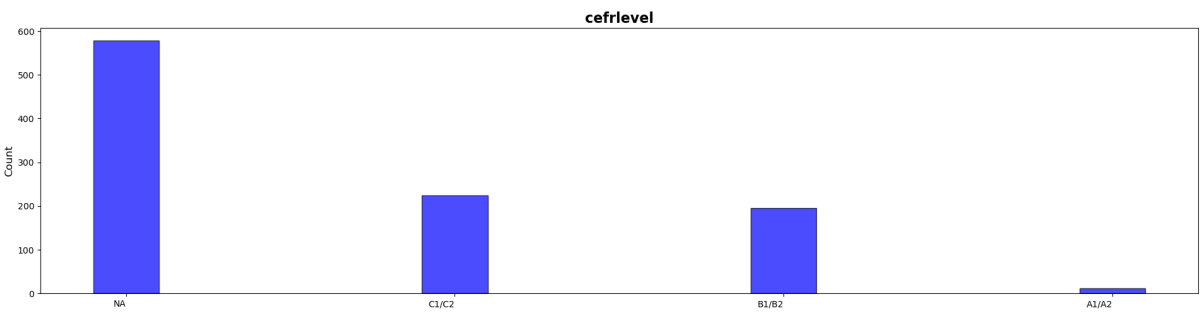

englishyears
Count

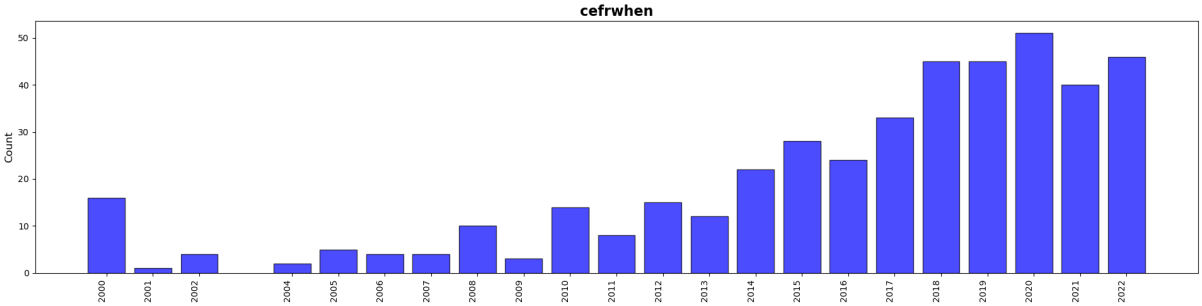

englishcountrymonths
Count
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 18 19 20 22 24 25 26 29 30 31 33 34 35 36

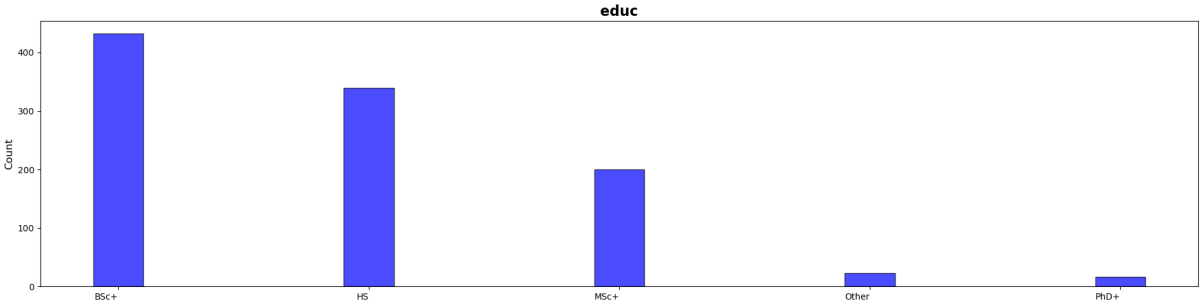

englishexposure
Count
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 23 24


otherlanguages
Count
Spanish German French Other Italian Russian Japanese Polish Afrikaans Portuguese Hindi Swedish Chinese Turkish Galician Dutch Catalan Arabic Croatian Zulu Basque Norwegian Korean Vietnamese Hungarian Greek Hebrew Bulgarian Slovak Xhosa Macedonian Tagalog Ukrainian Latvian Kannada Slovene Malay Finnish Urdu Punjabi Lithuanian


cefrself
Count
C1/C2 B1/B2 A1/A2


cefrlevel
Count
NA C1/C2 B1/B2 A1/A2


cefrwhen
Count
0
10
20
30
40
50
2000
2001
2002
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022


educ
Count
0
100
200
300
400
BSc+
HS
MSc+
Other
PhD+

A.4. Comparison of keystroke log data between tasks

For each participant, we compared the key press latency (i.e., the time interval between two consecutive key down events) t_{PL} for the copy-text task and the essay-writing task with an Anderson-Darling test. When the p-value of this test is high, there is evidence that the cognitive effort of the participant for the essay-writing task is not much different than for the copy-text task. This could potentially imply that the participants were transcribing text from other sources. Figure 4 shows the sorted p-values over 1,006 submissions.

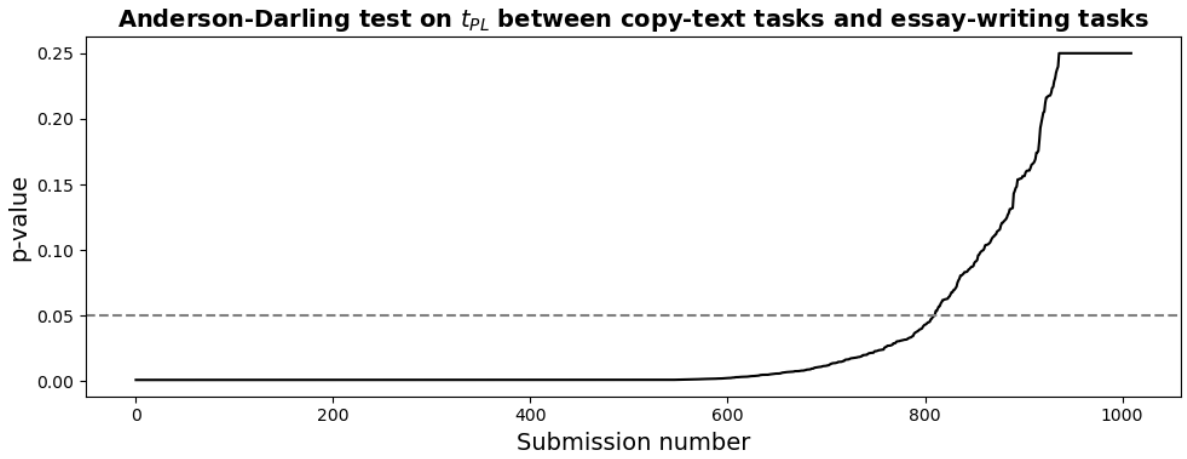


Figure 4: Anderson-Darling test p-values between the key press latency samples between the copy-text task and the essay-writing task of each participant. Submissions for which the p-value are large may imply similar cognitive efforts between the two tasks.