Joo Hyun Lee

Cornell University

Keystroke Logging in Second Language Writing Research

RWTH Aachen University
Department of English Linguistics
Elma Kerz, Marcus Ströbel

UROP – Undergraduate Research Opportunities Program

Aachen
July 12, 2019

Table of Contents:

Abstract

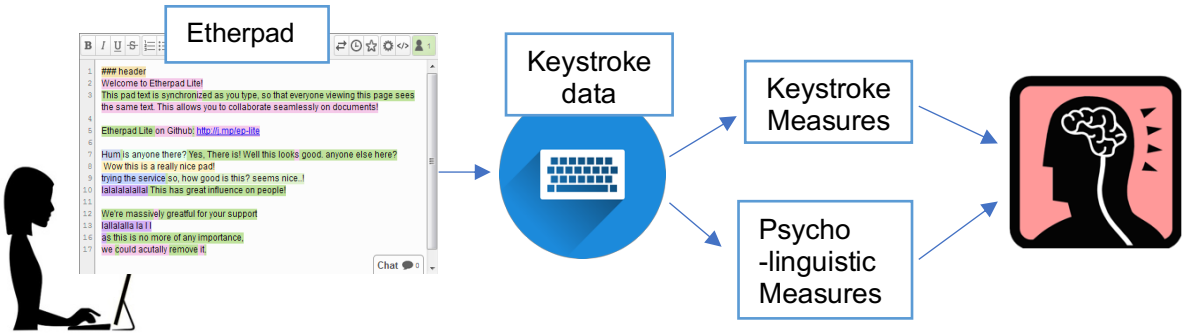
Many researchers in different fields have been interested in the cognitive process while writing. Recently, the development of keystroke logging techniques has provided new insight in understanding the cognitive process while writing. Using appropriate measures such as pause length between keystrokes, it is possible to infer the cognitive effort put in a specific part of a text. With large keystroke logging data from students' texts, we examined keystroke measures like pause lengths, burst lengths as well as psycholinguistic measures such as word frequency and n-gram frequency. The relationship between two different types of measures were examined. Many of the results support the usage-based and chunk-based writing process.

1. Introduction

Understanding human's key literacy skill has been an important but a challenging topic in many different fields such as linguistics, education, and cognitive science. Many researchers have been particularly interested in finding the cognitive process lying behind the writing process. As keyboard writing emerged recently, keystroke logging, which is a technique to record the timestamp and the keys pressed, has provided a new insight in understanding the cognitive process in writing. Several researches suggest meaningful measures that can be extracted from keystroke logging data. Commonly used measures are pauses between keystrokes, burst (consecutive keystrokes without pause), and revision (Alves, 2008; Baaijen, 2012; O'Brien, 2006; Miller, 2008).

Pauses and bursts are particularly helpful when revealing the writing process that actually happens. There are three phases in keyboard writing which are planning, translating, and revising. Pauses happen when the writer is in the planning or revising phase. (Alves, 2008) Bursts happen when the writer is translating phase, where the writer translates the idea into concrete words and types them. Knowing this, the corresponding writing phase at each time point can be reverse engineered using the pauses and bursts.

2. Project Description



We expect to see how cognitive efforts are embedded in the keystroke logging data. To achieve this, there are several challenges. First of all, the large amount of data must be cleaned and turned into a form that is easily manipulatable. Therefore, the Etherpad changeset needs to be understood. Second, appropriate keystroke measures that can be applied to the database must be chosen, whether they are already developed measures or new customized measures. Third, psycholinguistic

measures that represent the cognitive effort that is achievable from the data should be found. Lastly, the relationship between the keystroke measures and the psycholinguistic measures needs to be found.

Word frequency and n-gram frequency are important psycholinguistics measures, which represent how much the writer is the familiar with the word or the word sequence. Previous studies support the evidence of usage-based language acquisition and processing (Tomasello, 2009; Goldberg, 2006) Word frequency and keystroke measures can be compared to see if this usage-based language processing is also present in keyboard writing. Moreover, previous studies show that human process language by chunk (multiword sequence) due to limits in working memory. (Simon, 1974; McCauley, 2019) N-gram frequency and keystroke measures can be used to see if frequent n-grams (multiword sequences) require less cognitive effort.

3. Project Data / Conducted Research

3.1 Data

660 students wrote around 5 to 10 texts, each summarizing a linguistics lecture content, over a semester. We had an Etherpad change sets of text files approximately 3 GB in size in total. For each change set file, there was a corresponding text file the student wrote. Each Etherpad change set contained information of each keystroke. For example, the below change set represents the insertion of letter 'L' at time 1524302714328 (in milliseconds).


```
{"changeset":"Z:1>1*0+1$L","meta":{"author":"a.19M1ZmtK1EevYHcx","timestamp":152430
2714348}}
```


3.2 Preprocessing

Since the change sets were not easy to comprehend and accessible, only the necessary information was extracted and converted into CSV form. Pandas dataframe called 'keystrokes' was created to easily access the information. The position of the keystrokes in the final text was calculated by keeping track of the current position for every keystroke. When deletion occurred, all of the position after that keystroke would be decremented by 1. Moreover, to easily find the corresponding sentence for each keystroke, the final texts were parsed into sentences and numbered. Stanford CoreNLP was used to parse the sentences. These sentences were put in a separate pandas dataframe called 'sentences'. Based on position of keystrokes, sentence id was mapped to each keystroke. Using the same method, the texts were tokenized in words and put in a new dataframe called 'words'. word id was mapped to each key stroke. file id and user id (writer) were assigned to each keystroke using the file path.

Keystrokes Dataframe

	w_id	s_id	f_id	u_id	char	pos	t	end_t	op	path
0	0.0	0	0	0	L	0	2783551	1524218645504	+	SS18/01894/1
1	0.0	0	0	0	e	1	255	1524218645759	+	SS18/01894/1
2	0.0	0	0	0	a	2	255	1524218646014	+	SS18/01894/1
3	0.0	0	0	0	r	3	128	1524218646142	+	SS18/01894/1
4	0.0	0	0	0	n	4	129	1524218646271	+	SS18/01894/1

Words Dataframe

id	s_id	f_id	u_id	text	start_pos	end_pos	path
0	0	0	0	Learning	0	7	SS18/01894/1
1	0	0	0	Journal	8	15	SS18/01894/1
2	0	0	0	The	19	22	SS18/01894/1
3	0	0	0	second	23	29	SS18/01894/1
4	0	0	0	session	30	37	SS18/01894/1

Sentences Dataframe

	id	f_id	u_id	text	start_pos	end_pos	text_len	word_count	t
0	0	0	0.0	Learning Journal 1\n\nThe second session dealt...	0	71	71.0	10.0	2811019.0
1	1	0	0.0	In the following I will describe the most impo...	72	142	70.0	12.0	59341.0

In the keystroke data, there were some part of texts that were not appropriate for keystroke analysis. There were many copy-pasted parts, sentences that were too short or too long due to parsing problems, and parts with bullet points/lists. These were removed. The properties that were found later on from this given information from the data were added to columns in the appropriate dataframes.

w_id	word id
s_id	Sentence id
F_id	File id
U_id	User id
Char	character inserted, NaN if operation is deletion
pos	position of the keystroke in the final text
t	time from previous operation to current operation in milliseconds
end_t	time the current operation occurred
op	operation (+ : insertion, - : deletion
path	substring of file path
start_pos	Starting character position of word or sentence
end_pos	ending character position of word or sentence
text_len	Number of characters
word_count	Number of words

3.3 Measures

We first explored many different properties of keystroke data, such as pause length, pause distribution, deletion, etc. as well as different psycholinguistics measures. Among different keystroke measures, pause length was a reliable and relatable measure to word frequency and n-gram frequency because the amount pause before a word represents the amount of cognitive effort put in each word. However, there were some outliers, so we used t_filter_1000 or median of pause length when necessary. Also, we used burst length to examine if greater pause length leads to larger burst length. within_burst was used to distinguish n-grams across burst and within burst to observe if there was difference in n-gram frequency.

3.3.1 Keystroke Measures

Measure	Name	Description
t	Pause length	Pause length between keystrokes/words
t_filter_10000	Pause length filtered (10000ms)	Maximum of pause length between keystrokes/words and 10000ms
avg_burst _len	Average Burst Length	Average burst length in a file
within_burst	Within/across burst	Binary number (0,1) weather the n-gram is within a burst

3.3.2 Psycholinguistics Measures

Measure	Description
Word frequency	Word frequency in scale of 0 to 1
n-gram frequency	N-gram frequency in scale of 0 to 1
Occurrence in data	The number of occurrence of words in data

3.4 Methods

3.4.1 Visualization of Burst

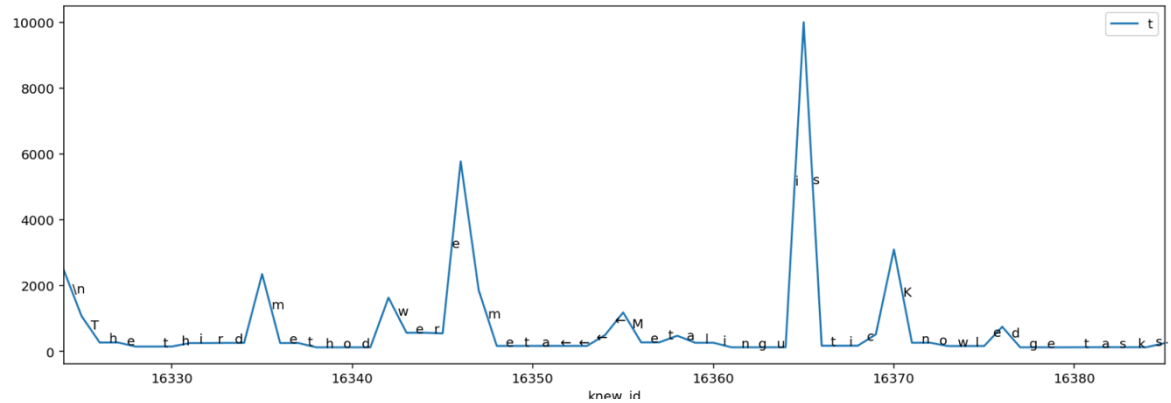
A function to retrieve and visualize the bursts and pauses was written to compare length and content of bursts depending on different pause length thresholds. Two versions of functions were made: (1) consider all pauses including within words (2) only consider pauses between words. Below is an example output of the first function with three different pause thresholds 550ms, 1000ms, 2000ms. The numbers inside curly bracket shows the pause lengths greater than the given threshold. Left arrow represents backspace.

```
 {0}Learne←←←←←← {2298}The term Learner Corpus comprises a cin {574}←←oncept...\n {635}-
{691}←←deisg {776}←←←sign considerations for compiling lean←rner corpora {2452}\na {1280}
←example of a nwe←←ew, cm←ompiled {639} learner corpus {1199} sre←←←resource to ills←ust
rate u←such cos←nsideratiosn←←ns\n\n {124925}← {124925}← {124925}← {124926}← {124925}←
{124925}← {124926}← {639}collection {1493} of texts that are being produced  {760}by {1436}
```

```
 {0}Learne←←←←←← {2298}The term Learner Corpus comprises a cin←←oncept...\n- ←←deisg←
←←sign considerations for compiling lean←rner corpora {2452}\na {1280}←example of a nwe←←
ew, cm←ompiled learner corpus {1199} sre←←←resource to ills←ustrate u←such cos←nsiderati
osn←←ns\n\n {124925}← {124925}← {124925}← {124926}← {124925}← {124925}← {124926}←collec
```

```
 {0}Learne←←←←←← {2298}The term Learner Corpus comprises a cin←←oncept...\n- ←←deisg←
←←sign considerations for compiling lean←rner corpora {2452}\na←example of a nwe←←ew, cm
←ompiled learner corpus sre←←←resource to ills←ustrate u←such cos←nsideratiosn←←ns\n\n
{124925}← {124925}← {124925}← {124926}← {124925}← {124925}← {124926}←collection of texts
```

A graph was also helpful in observing the general distribution of the pauses and bursts.



3.4.2 Word Frequency and Pause Length

Before bigram/trigram frequency and bursts were compared, we wanted to examine the relationship between word frequency and pause length. Word frequency (from COCA - academic) and pause length were compared. Only the words with same length and no spelling errors were compared. Also, the number of occurrence of words in the data and pause length were compared.

3.4.3 N-gram Frequency and Bursts

First, the list of bursts using different pause thresholds was retrieved. Appropriate pause thresholds (2*median of pause threshold, 550ms, 1000ms, and 2000ms) were chosen by looking into the burst visualization and also referring to related papers. (Rosenqvist, 2015) Then, the n-gram frequency (in log scale) within a burst and across bursts were compared. For instance, consider the sentence '*The second aspect is time constraint.*' Assume the bursts are 'The second aspect is' and 'time constraint.' for pause threshold 550ms. Comparing the bigram frequency within bursts and across bursts would take following steps. For within burst, the average bigram frequency (in log scale) of {The second, second aspect, aspect is, time constraint} would be taken. Then, for across bursts, the average bigram frequency (in log scale) of {is time} would be taken. These steps were executed for all the

sentences and different pause thresholds. The histogram and the means of two groups (within burst, across bursts) were observed.

3.4.4 Time Between Keystrokes

The time between keystrokes were categorized into five different types shown as below. _ is space. * shows the where the time interval is. a*b represents time between keystroke 'a' and keystroke 'b'.

Type	Description	Example
0	Within word (letter to letter)	h*ello
1	Last letter of word to space	I*_am
2	Space to first letter of next word	I_*am
3	Period (end of sentence) to space	here.*_You
4	Space to letter (beginning of sentence)	here._*You

3.4.5 Average Burst Length and Pause Length

The average burst length and the median pause length between all the bursts for each file were calculated and graphed in 10-quartile boxplots since the distribution was not even. Median pause length was chosen because of the outliers.

4 Project Result

4.1 Descriptive Statistics

After cleaning up the data, 1/3 of the original data remained as shown in the table. Additional information about the data is also shown below.

Data	Number
Writers	660
Files	7,012
Sentences	
Word	
Keystrokes	
Size	~3 GB

Clean up

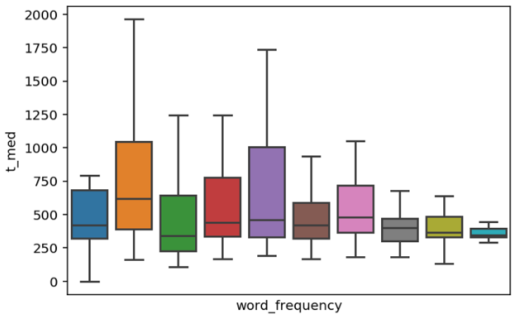
Data	Number
Writers	513
Files	3,466
Sentences	95,354
Word	1,831,361
Keystrokes	14,966,923
Size	~1 GB

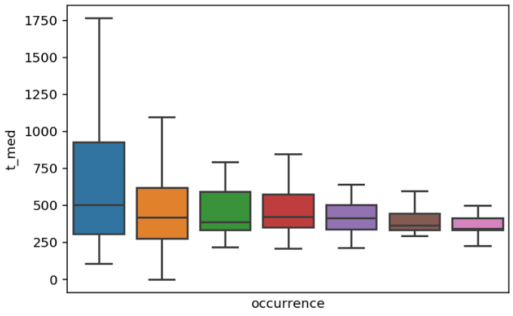
Statistics	Value
Word Per Minute (t_filter_2000)	24.66
Word Per Minute (t_filter_10000)	18.56
Average Sentence Length (characters)	118.46
Average Sentence Length (words)	19.21
Average Word Count in each text file	530.41
Average time spent on each character (t_filter_10000)	379.30ms
Median time spent on each character (t_filter_10000)	174ms
Average time spent on each character (t_filter_2000)	274.34ms
Median time spent on each character (t_filter_2000)	172ms

4.2 Word Frequency and Pause Length

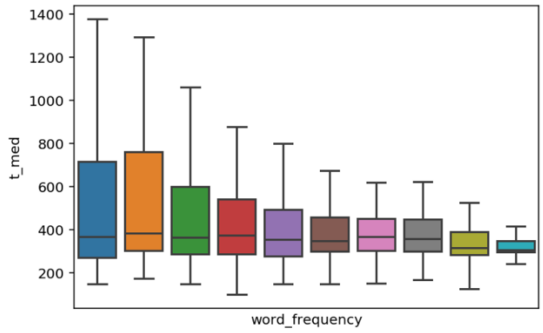
As shown below, generally the median pause length decreased when the word frequency was higher. The effect was less evident when the word length was small, especially when smaller than 4. Below are the correlation coefficient matrix and the 10-quantile box plots of x = word frequency/occurrence and y = median time.

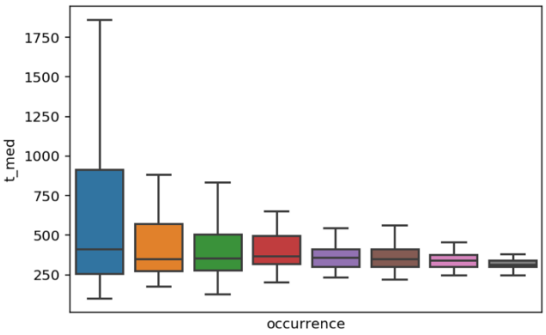
Word length = 3	Mean time	Median time	Occurrence
Word frequency	-0.014170	-0.057664	0.980847
Occurrence	-0.011448	-0.046947	1



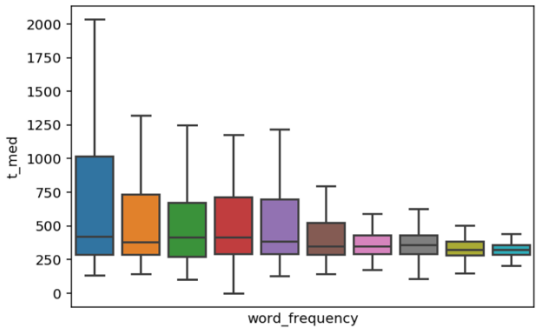


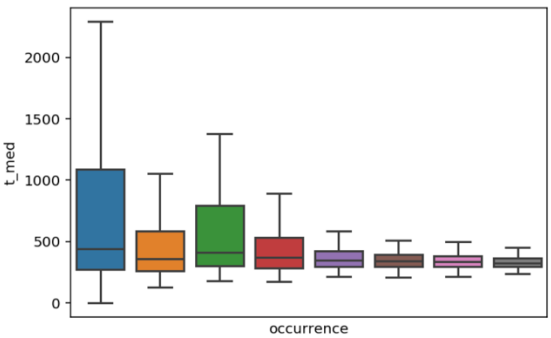
Word length = 5	Mean time	Median time	Occurrence
Word frequency	-0.038743	-0.086543	0.728264
Occurrence	-0.030937	-0.066766	1





Word length = 7	Mean time	Median time	Occurrence
Word frequency	-0.089100	-0.135862	0.438591
Occurrence	-0.063493	-0.098984	1





4.3 N-gram Frequency and Bursts

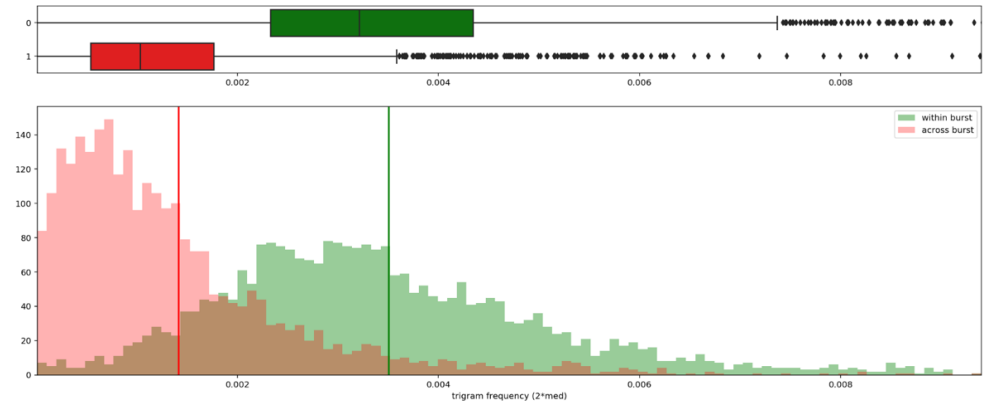
In all of the cases, the n-gram frequency was significantly higher for the n-grams within burst (green in graph) than across burst (red in graph). P-values were calculated in each case where the null hypothesis is H_0: $\mu_0 = \mu_1$. (μ_0:mean of bigram frequency within burst, μ_1: mean of bigram frequency across bursts)

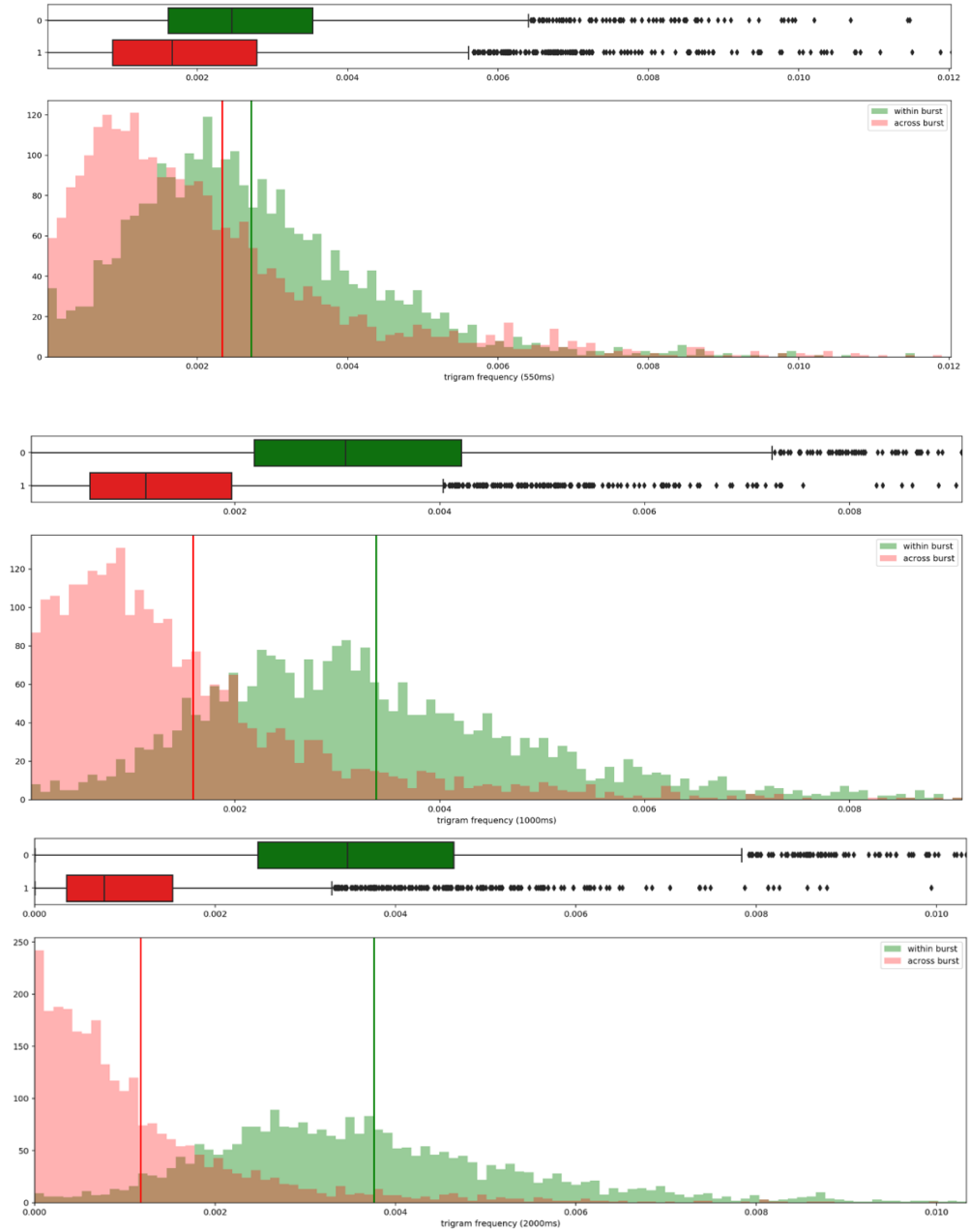
Bigram Frequency with different pause threshold (2*pause length median, 550ms, 1000ms, 2000ms)

Pause Threshold	T-statistics	P-Value
2*median	36.6999096887	1.29972404367e-263
550ms	44.2518135482	≅0.0
1000ms	39.055509109	3.74521469014e-293
2000ms	35.1555921505	9.57787712242e-244

Trigram Frequency with different pause threshold (2*pause length median, 550ms, 1000ms, 2000ms)

Pause Threshold	T-statistics	P-Value
2*median	40.6564082057	≅0.0
550ms	5.56547682304	2.78747796899e-08
1000ms	30.8025658713	1.18514973243e-191
2000ms	56.2489479571	≅0.0





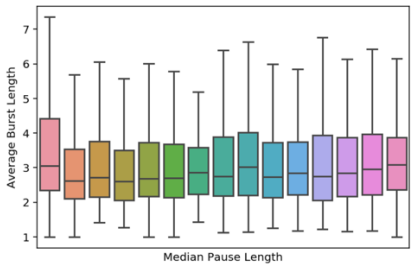
4.4 Time between Different keystrokes

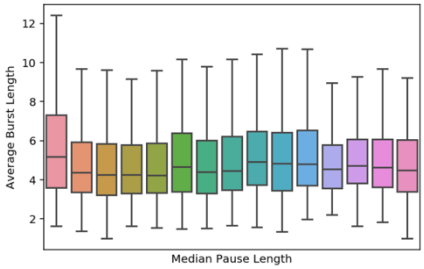
The time (pause) taken between different types of keystrokes were compared. Both mean and median were high in order of between sentences, between words, within words. This shows that people spend more time right before starting the sentence or a new word.

keyst_type	count	description	mean	median
-1.0	732823	not classified	704.277587	258
0.0	8983569	within words	283.525927	170
1.0	2022161	between words(last char)	375.577840	174
2.0	2743465	between words(space)	557.161356	249
3.0	111546	between sentences(period)	780.779069	495
4.0	206103	between sentences(space)	849.857595	252

4.5 Average Burst Length and Pause Length

Average burst length and median pause length between bursts in each file were compared. Below graphs are the results when the pause thresholds were 550ms and 1000ms. The correlation coefficient is shown in the table. As median pause length increased, average burst length increased. This shows that people pause a longer period before writing a greater length of bursts. However, in 2000ms case, this effect was not observed.





Pause Threshold	Correlation Coefficient
550ms	0.412974
1000ms	0.096404
2000ms	-0.006937

5 Evaluation

5.1 Conclusion

As expected, as word frequency increased, the time taken to type the words decreased. However, it was interesting that this effect is less clear when the length of the word is short. This may imply that when the word length is short, the writer does not take additional time to think about and write a more unfamiliar word. The relationship between n-gram frequency and the pause length show that people generally type without pausing when writing more frequent bigrams and trigrams. Both findings were meaningful because we could see that the chunk-based process also happens in keyboard writing.

There were also some shortcomings in this research. Because the Etherpad recorded every 500ms, we had to estimate the time for faster keystrokes. Also, we could not distinguish the pauses that were due to incompetence in typing and pauses that were needed for translation to language from idea. A control group where the same writers type a given script may be helpful in a similar future research.

5.2 What I learned

Working on this research, I have learned how to interpret Etherpad changesets and work with large database. I also experienced working with many python libraries that are crucial in natural language processing and graphing such as nltk, pandas, numpy, seaborn, etc. Our code had to be efficient since we were dealing with large data. So, I have learned many different methods to use pandas dataframe in a more efficient way.

Moreover, we worked as a team on interpreting the data and pre-processing. Through collaborating with other people, I have learned how helpful it is to share different ideas and insights to understand a big data set. Having a common dataframe we could work on, we could easily discuss about different ideas and experiment on our own. We branched out to looking at different units of the keystrokes and at the end we all could come up with interesting results.

Reference

1. Baaijen, Veerle M., et al. “Keystroke Analysis.” *Written Communication*, vol. 29, no. 3, 2012, pp. 246–277., doi:10.1177/0741088312451108.
2. Chan, Sathena. “Using Keystroke Logging to Understand Writers' Processes on a Reading-into-Writing Test.” *SpringerLink*, Springer International Publishing, 28 June 2017, link.springer.com/article/10.1186/s40468-017-0040-5.
3. Lacruz, Isabel, et al. *Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing*. 2012.
4. Leijten, Mariëlle, and Luuk Van Waes. *Keystroke Logging in Writing Research: Using Input Log to Analyze and Visualize Writing Process*. 2013.
5. O'Brien, Sharon. *Pauses as Indicators of Cognitive Effort in Post-Editing Machine Translation Output*. 2006, doras.dcu.ie/17156/1/01_1_21ps_Across_Pauses.pdf.
6. Miller, Kristyan Spelman, et al. “The Psycholinguistic Dimension in Second Language Writing: Opportunities for Research and Pedagogy Using Computer Keystroke Logging - MILLER - 2008 - TESOL Quarterly - Wiley Online Library.” *TESOL Quarterly*, John Wiley & Sons, Ltd, 30 Dec. 2011, onlinelibrary.wiley.com/doi/abs/10.1002/j.1545-7249.2008.tb00140.x.
7. Wengelin, Åsa. *Combined Eye Tracking and Keystroke-Logging Methods for Studying Cognitive Processes in Text Production*. 2009.
8. Simon, Herbert A. “How Big Is a Chunk?” *Science*, American Association for the Advancement of Science, 8 Feb. 1974, science.sciencemag.org/content/183/4124/482.
9. Tomasello, Michael, and Patricia J. Brooks. “Young Children's Earliest Transitive and Intransitive Constructions.” *Cognitive Linguistics (Includes Cognitive Linguistic Bibliography)*, Walter De Gruyter, Berlin / New York, 5 Aug. 2013, www.degruyter.com/view/j/cogl.1998.9.issue-4/cogl.1998.9.4.379/cogl.1998.9.4.379.xml.
10. Ivic, Milka. “Adele E. Goldberg, Constructions at Work, the Nature of Generalization in Language, Oxford 2006, Oxford University Press.” *Juznoslovenski Filolog*, no. 62, 2006, pp. 367–368., doi:10.2298/jfi0662368i.
11. Alves, Rui Alexandre, et al. “Execution and Pauses in Writing Narratives: Processing Time, Cognitive Effort and Typing Skill.” *International Journal of Psychology*, vol. 43, no. 6, 2008, pp. 969–979., doi:10.1080/00207590701398951.
12. Ströbel, Marcus, et al. *CoCoGen - Complexity Contour Generator: Automatic ...*2016, www.aclweb.org/anthology/W16-4103.
13. Rosenqvist, Simon. “Developing Pause Thresholds for Keystroke Logging Analysis.” 2015.
14. McCauley, Stewart, and Morten Christiansen. *Language Learning as Language Use: A Cross-Linguistic ...*2019, cnl.psych.cornell.edu/pubs/2019-mc-PsychRev.pdf.