Institut für Anglistik, Amerikanistik und Romanistik

RWTH AACHEN UNIVERSITY

Keystroke Logging in Second Language Writing Research

Joo Hyun Lee, Elma Kerz & Marcus Ströbel

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 has provided a new insight in understanding the cognitive process in writing. Commonly used measures are pauses between keystrokes, burst (consecutive keystrokes without pause), and revision (Alves, 2008; Baaijen, 2012; O'Brien, 2006; Miller, 2008).

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 to be cleaned and turned into a form that is easily manipulatable. Second, appropriate keystroke measures that can be applied to the database must be chosen. Third, psycholinguistic measures that represents 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.

Discussion

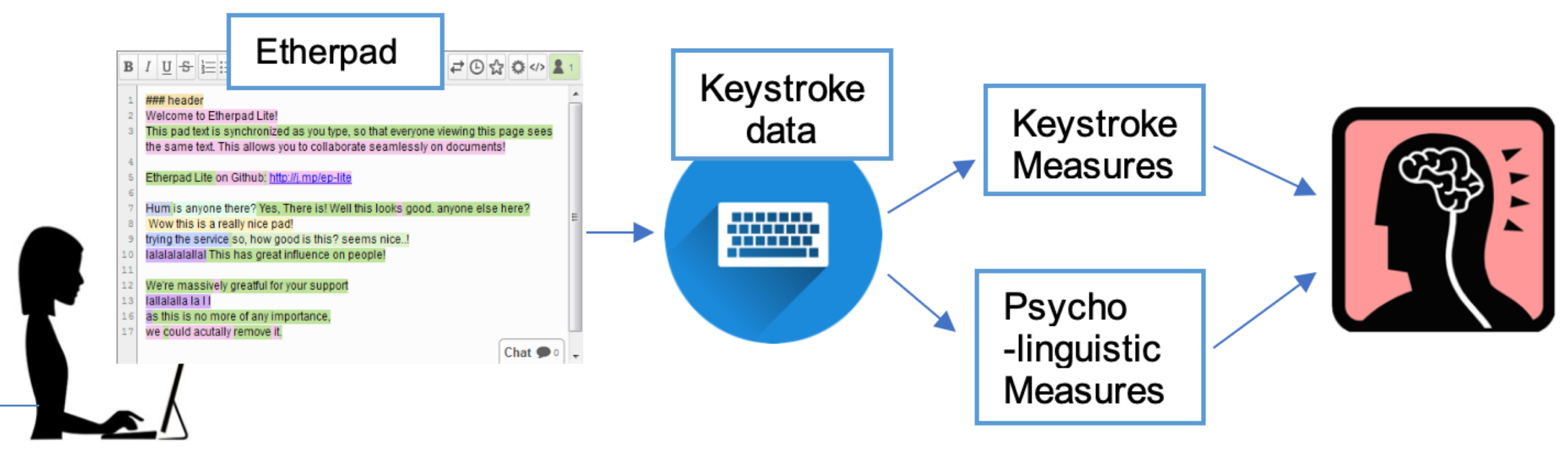
As expected, as word frequency increased the time taken to type the words decreased. However, it was interesting that this effect is less clear when the length of the word is short. This may imply that when the word length is short, the writer does not take additional time to think about and write a 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.

Methods

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 the corresponding text file the student wrote. Each Etherpad change set contained information of each keystroke. 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. Words, sentences, files, users dataframes were also created.

Keystroke Measure	Name
t	Pause length
t_filiter_10000	Pause length filtered (10000ms)
avg_burst _len	Average Burst Length
within_burst	Within/across burst

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



Schematics of the approach

Word Frequency

Before bigram/trigram frequency and bursts were compared, we examined 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.

N-gram Frequency and Bursts

First, the list of bursts using the different pause thresholds were 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. Then, the n-gram frequency (in log scale) within a burst and across bursts were compared. 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.

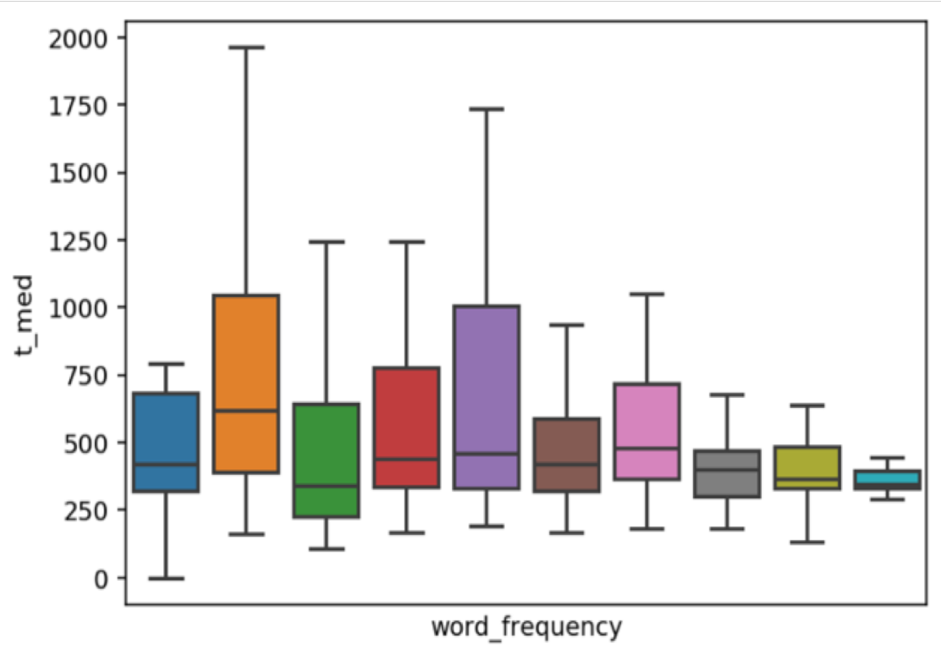
Results

Statistics

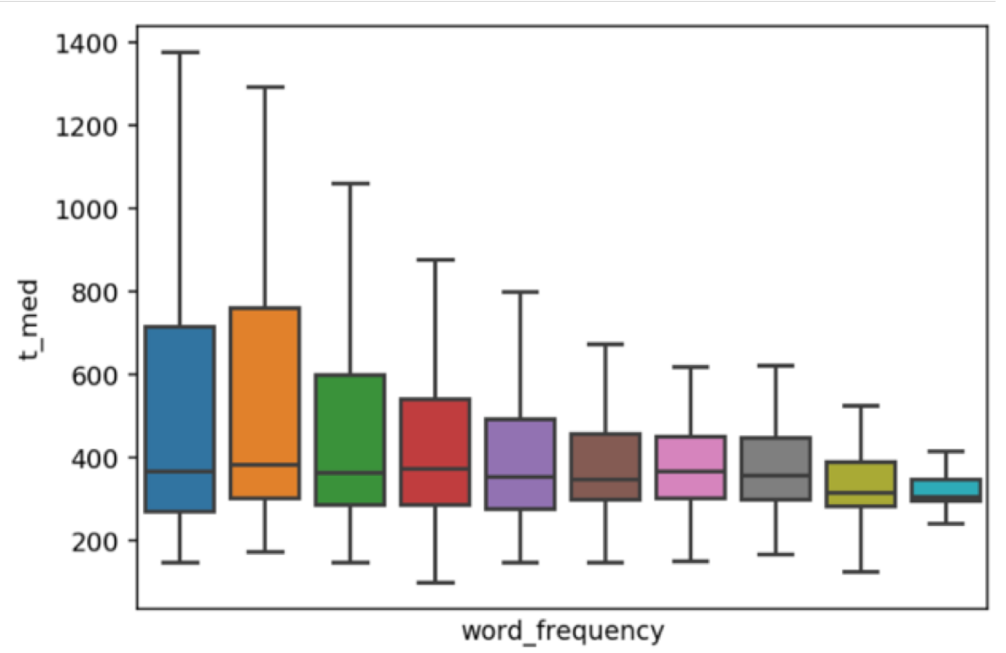
Statistics	Value
Word Per Minute (t_filter_2000)	24.66
Word Per Minute (t_filter_10000)	18.56
Average Sentence Length (characters)	118.46
Average Sentence Length (words)	19.21
Average Word Count in each text file	530.41
Average time spent on each character (t_filter_10000)	379.30ms
Median time spent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

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.



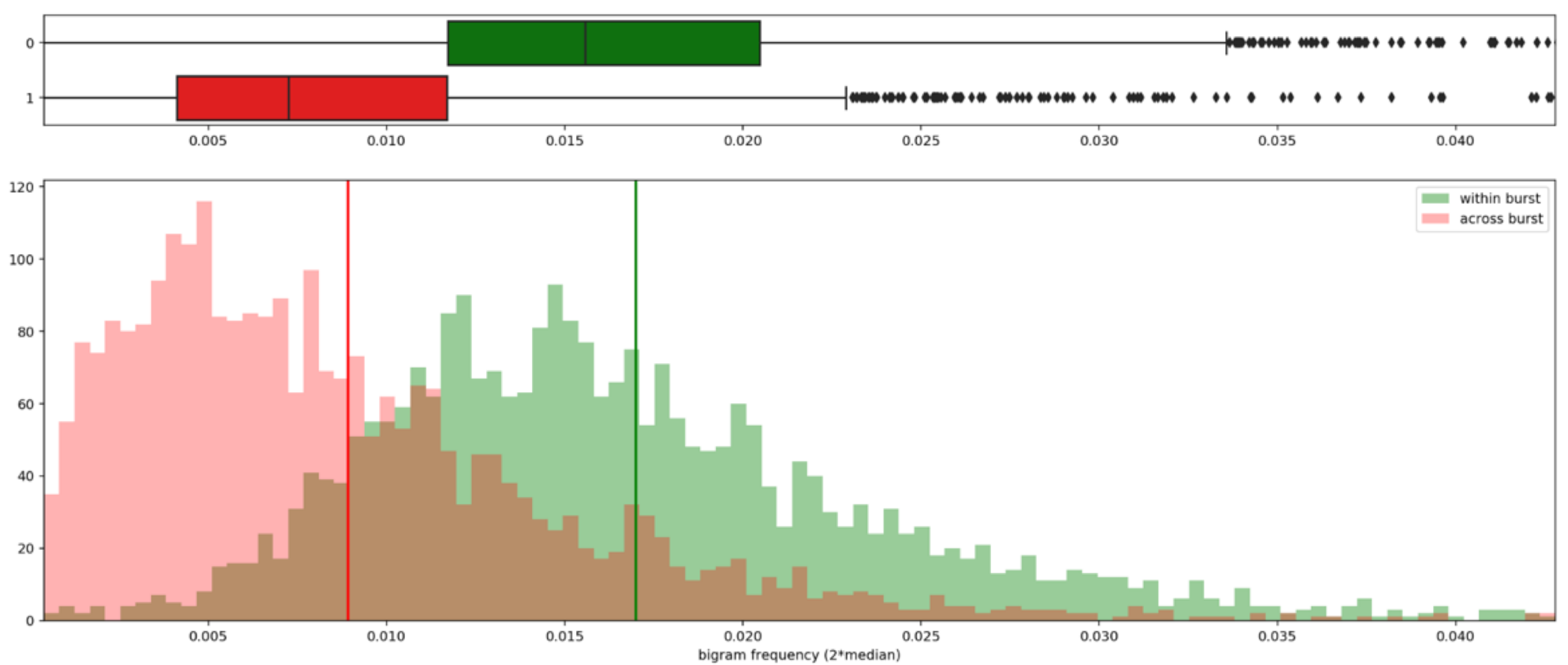
Word frequency and Median Pause Length (word length = 3)



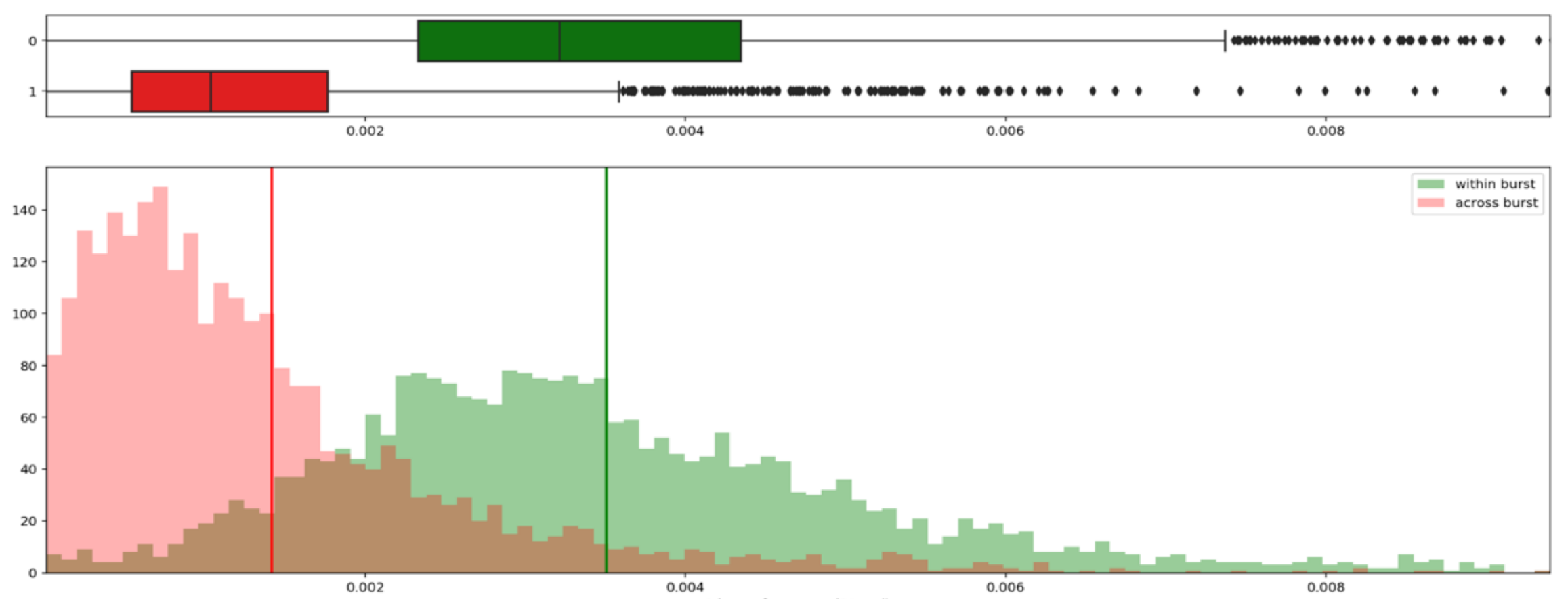
Word frequency and Median Pause Length (word length = 5)

N-gram Frequency and Bursts

In all of the cases, the n-gram frequency was significantly higher for the n-grams within burst (green) than across burst (red).

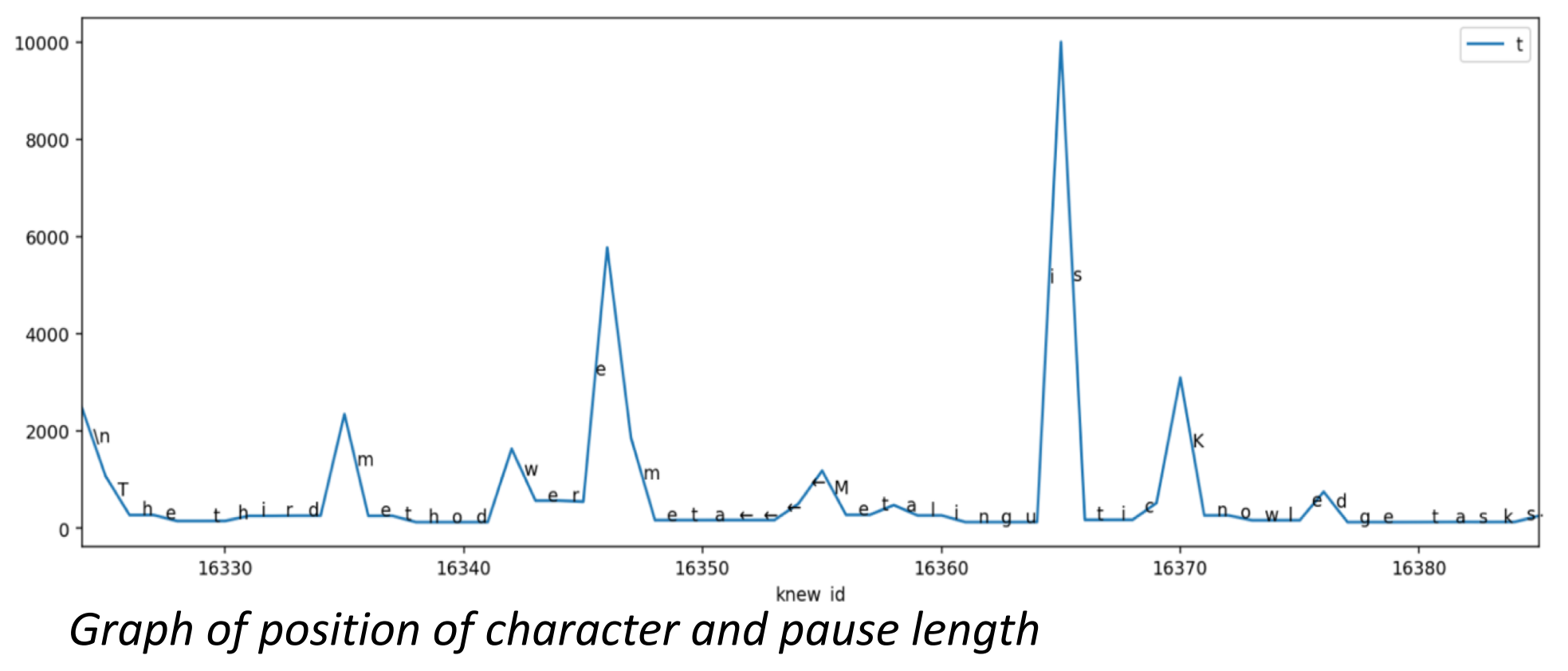


*Bigram Frequency (2*median pause length) within vs across burst*



*Trigram Frequency (2*median pause length) within vs across burst*

Visualization of Bursts



Graph of position of character and pause length

```
{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}
```

Different Keystroke Types

The time (pause) taken between different types of keystrokes were compared.

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