Institut für Anglistik, Amerikanistik und Romanistik

RWTH AACHEN UNIVERSITY

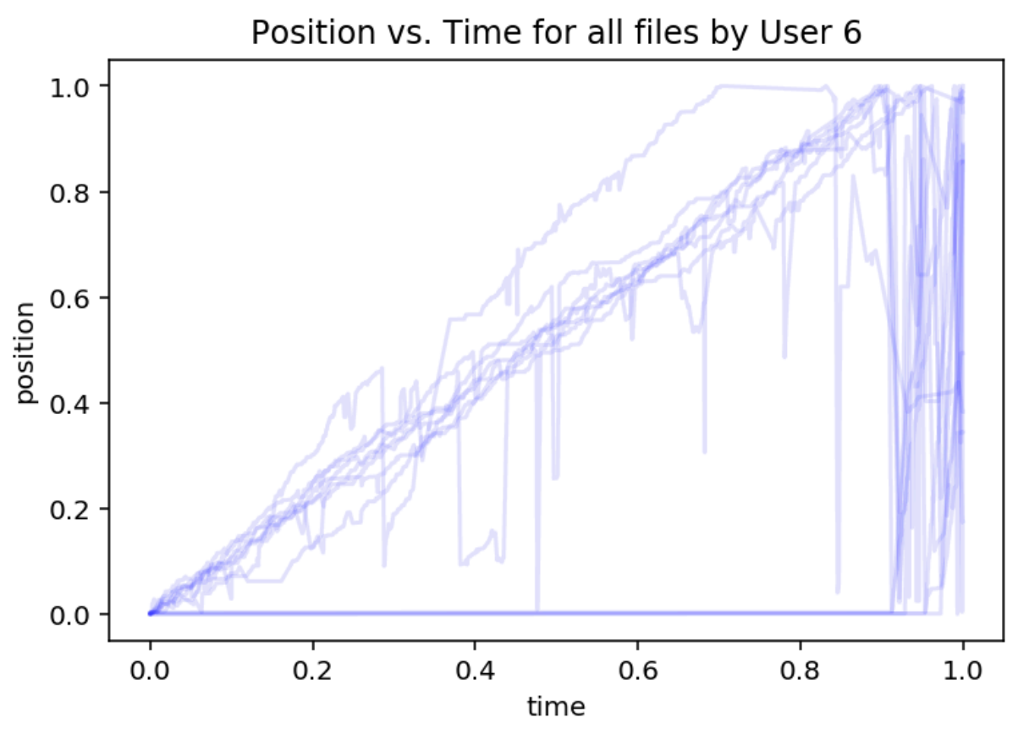
Keystroke Logging in Second Language Writing Research

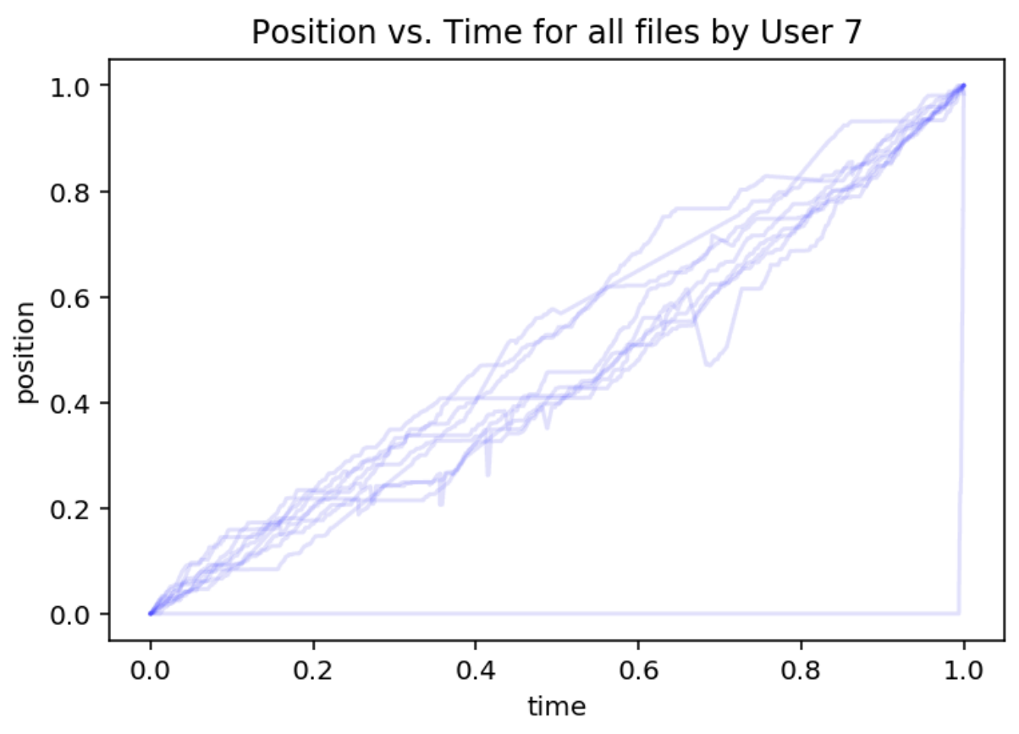
Logan Tidstrom, Elma Kerz, Marcus Ströbel

Abstract

The study of writing is an important tool to examine the underlying cognitive processes of language production. However, a lot of writing research focuses on written text as an end product. Conversely, keystroke logging provides data on the writing process as a text develops. Using the Etherpad text editor, we collected keystroke logging data from 660 university students, producing a total of 7012 files. The keystroke logging data was then compared with data regarding text complexity to analyze whether a correlation exists between the writing process and the complexity of a text.

Applications of Keystroke Logging





Keystroke logging reveals that User 6 tends to revise their text thoroughly once it is completed, whereas User 7 writes with very little revision at all. Keystroke logging can help us to identify and understand different writing processes.

Measure Extraction

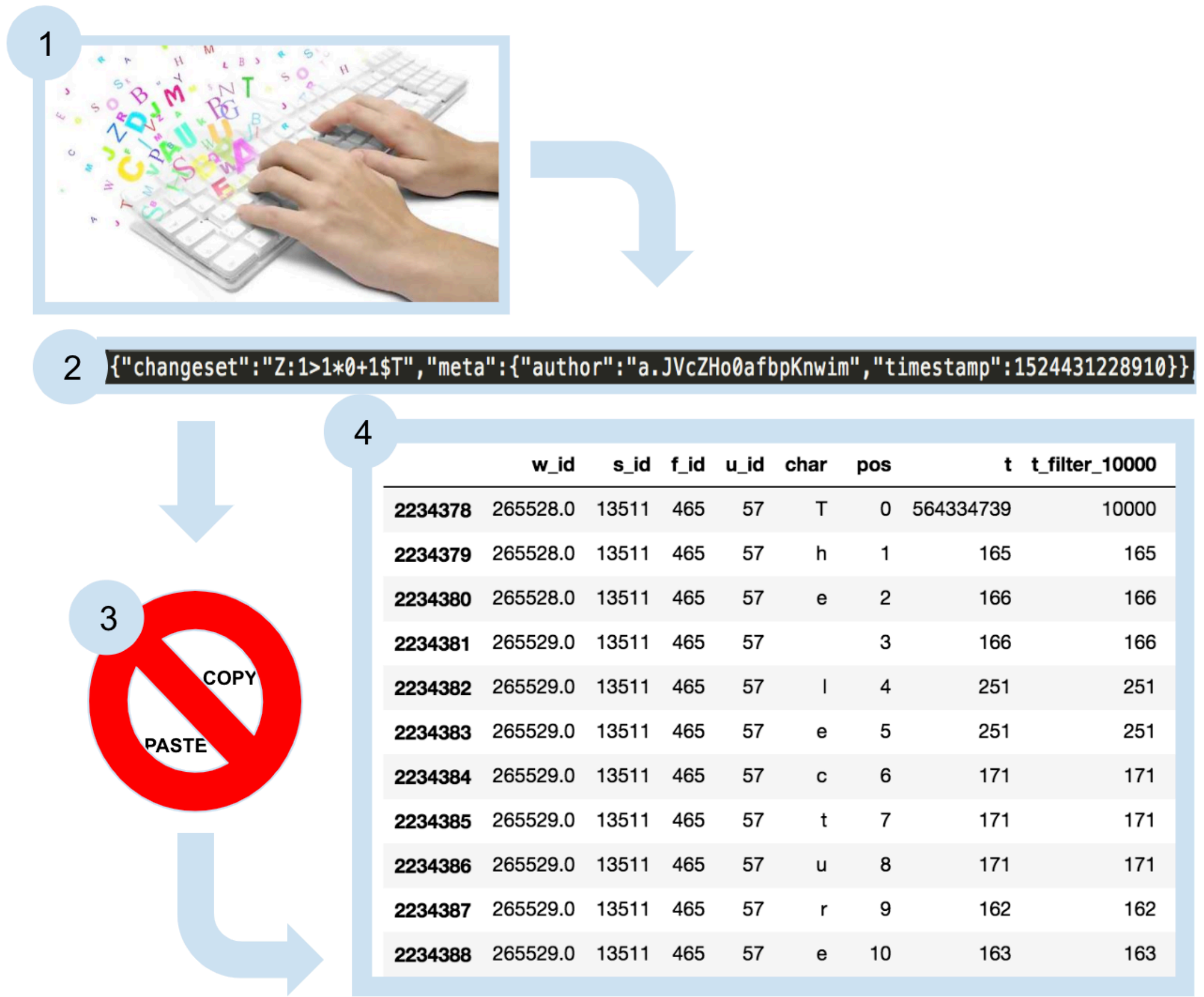
Keystroke Logging Measures

We split each of the texts into individual sentences using the Stanford CoreNLP software.

Using the keystroke data, we developed a number of measures pertaining to the manner in which each sentence was produced. The seven keystroke measures relevant to this study are defined below:

Keystroke Measure	Definition
word_count	Number of words in sentence
t_filter_10000	Total time spent composing the sentence, with all keystrokes over 10 seconds set to 10000ms
jumps	Number of times the writer jumped to non-consecutive positions within the sentence
chunks	Number of times the writer returned to the sentence from a different sentence in the file
word_t/ t_filter_10000	Percentage of total filtered sentence time that was spent within a word (i.e. sum of time spent on all *letter* keystrokes)
separator_t/ t_filter_10000	Percentage of total filtered sentence time that was spent on a space or new line character
revision_t/ t_filter_10000	Percentage of total filtered sentence time spent on a sentence excluding the first draft (i.e. excluding the time spent typing the first character at each position)

Data Preprocessing



```
{"changeset":"Z:1>1*0+1$T","meta":{"author":"a.JVcZHo0afbpKnwim","timestamp":1524431228910}}
```

	w_id	s_id	f_id	u_id	char	pos	t	t_filter_10000
2234378	265528.0	13511	465	57	T	0	564334739	10000
2234379	265528.0	13511	465	57	h	1	165	165
2234380	265528.0	13511	465	57	e	2	166	166
2234381	265529.0	13511	465	57		3	166	166
2234382	265529.0	13511	465	57	l	4	251	251
2234383	265529.0	13511	465	57	e	5	251	251
2234384	265529.0	13511	465	57	c	6	171	171
2234385	265529.0	13511	465	57	t	7	171	171
2234386	265529.0	13511	465	57	u	8	171	171
2234387	265529.0	13511	465	57	r	9	162	162
2234388	265529.0	13511	465	57	e	10	163	163

1. Etherpad collects keystroke logging data.
2. Changeset data is translated to reflect the time and type of each keystroke.
3. All copy/pasted text is removed from the data set, as it does not accurately reflect the writing process.
4. Keystroke data is stored in a Pandas data frame for analysis.

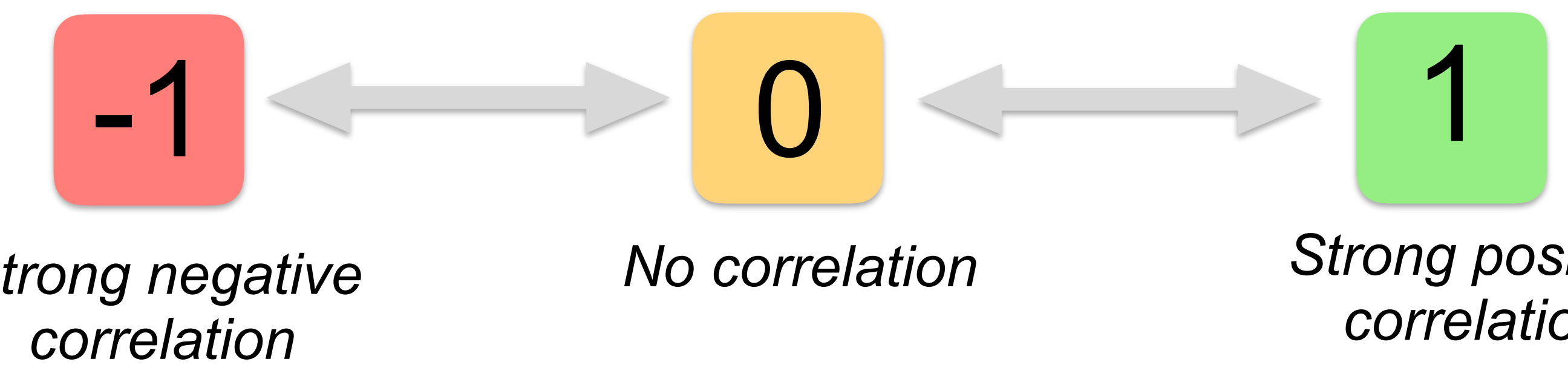
CoCoGen Measures

The Complexity Contour Generator (CoCoGen) provides lexical, syntactic, and morphological complexity measures for a text using a unique "sliding window technique". We ran CoCoGen on the split texts to obtain complexity scores for each sentence.

Results

To evaluate the correlation between keystroke and complexity measures, we computed the Pearson correlation coefficient for each pair.

Pearson Correlation Coefficients



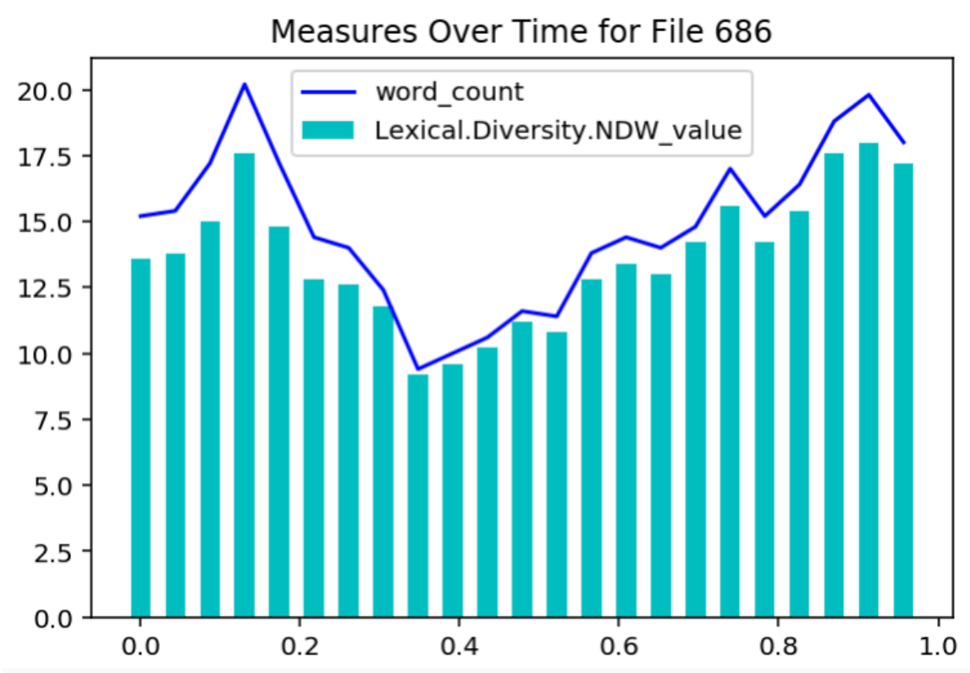
Strong negative correlation

No correlation

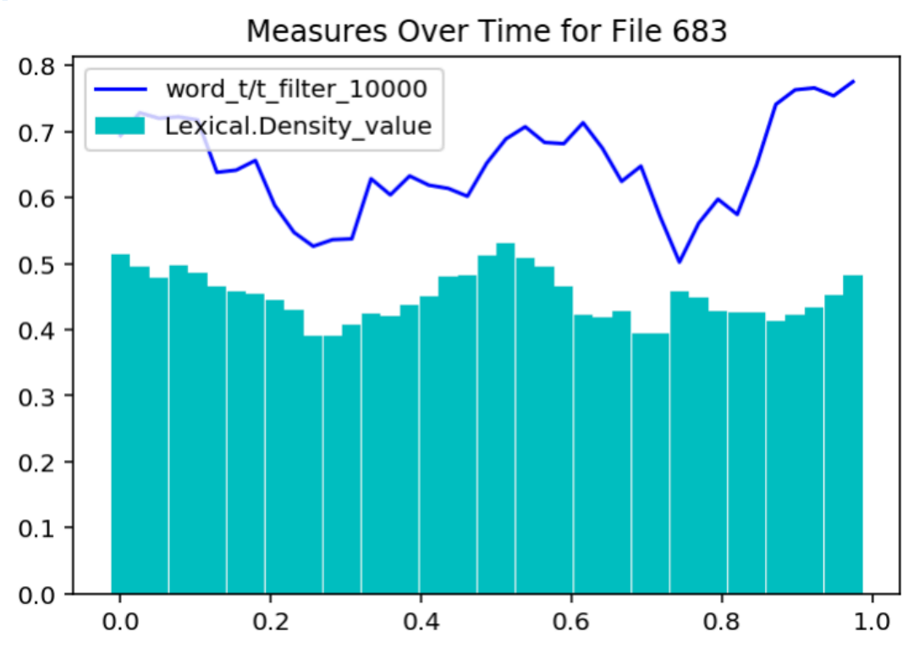
Strong positive correlation

	word_count	t_filter_10000	jumps	chunks	word_t/t_filter_10000	separator_t/t_filter_10000	revision_t/t_filter_10000
KolmogorovDeflate_value	0.88	0.67	0.29	0.17	0.05	0.12	0.04
Lexical.Density_value	-0.09	0.06	0.02	0.01	0.10	-0.12	-0.02
Lexical.Diversity.CNDW_value	-0.59	-0.39	-0.18	-0.10	0.03	-0.14	-0.04
Lexical.Diversity.CTTR_value	0.67	0.52	0.22	0.14	0.02	0.12	0.03
Lexical.Diversity.NDW_value	0.95	0.70	0.29	0.18	-0.00	0.17	0.05
Lexical.Diversity.RTTR_value	0.72	0.56	0.23	0.14	0.01	0.14	0.04
Lexical.Diversity.TTR_value	-0.59	-0.38	-0.17	-0.10	0.03	-0.14	-0.04
Lexical.Sophistication.AFL_value	0.14	0.08	0.03	0.02	0.01	0.01	0.00
Lexical.Sophistication.ANC_value	-0.11	0.06	0.04	0.02	0.03	-0.03	-0.02
Lexical.Sophistication.BNC_value	-0.19	-0.01	0.01	0.00	0.06	-0.09	-0.03
Lexical.Sophistication.NAWL_value	-0.01	0.06	0.02	0.01	0.04	-0.05	-0.00
Lexical.Sophistication.NGSL_value	-0.09	0.04	0.03	0.02	-0.01	0.02	-0.03
Morphological.KolmogorovDeflate_value	-0.84	-0.65	-0.27	-0.16	-0.07	-0.11	-0.04
Morphological.MeanLengthWord_value	-0.13	0.10	0.04	0.02	0.20	-0.23	-0.02
Morphological.MeanSyllablesPerWord_value	-0.11	0.10	0.04	0.03	0.16	-0.20	-0.02
NounPhrasePostModificationWords_value	0.36	0.32	0.16	0.11	-0.04	0.11	0.02
NounPhrasePreModificationWords_value	0.03	0.05	0.03	0.01	0.01	0.01	-0.01
Syntactic.ClausesPerSentence_value	0.59	0.39	0.17	0.10	-0.02	0.12	0.05
Syntactic.ComplexNominalsPerSentence_value	0.70	0.55	0.24	0.15	0.01	0.11	0.03
Syntactic.CoordinatePhrasesPerSentence_value	0.35	0.30	0.14	0.07	0.02	0.04	0.01
Syntactic.DependentClausesPerSentence_value	0.51	0.34	0.14	0.09	0.00	0.09	0.04
Syntactic.KolmogorovDeflate_value	-0.76	-0.59	-0.25	-0.15	-0.06	-0.10	-0.03
Syntactic.VerbPhrasesPerSentence_value	0.60	0.40	0.16	0.10	0.01	0.10	0.05

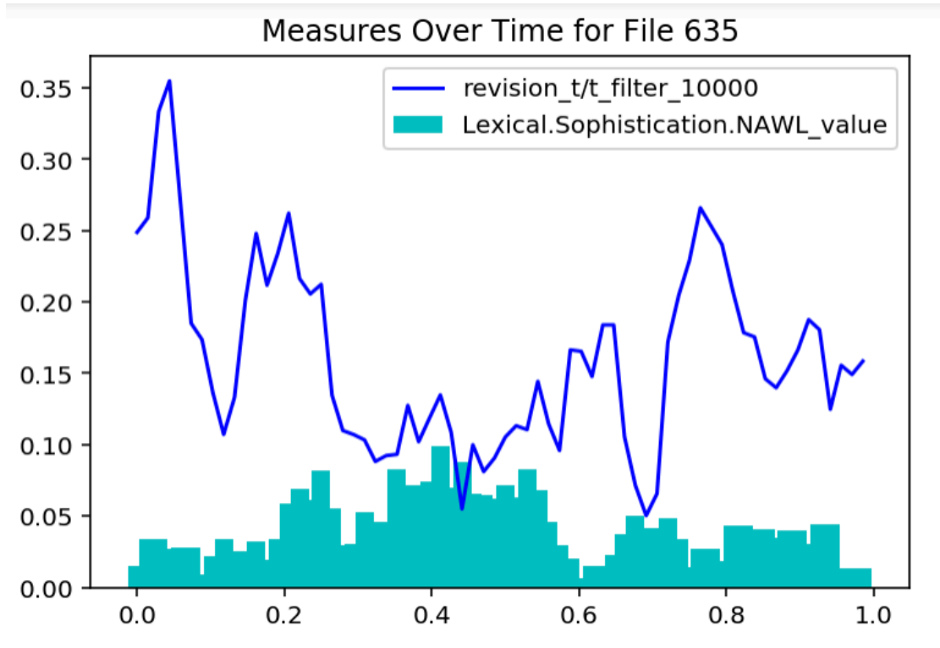
Pearson correlation coefficients for each pair of keystroke and complexity measures used in this study. Somewhat positive correlations (between 0.1 and 0.7) are highlighted in blue.



Word_count and Lexical.Diversity.NDW_value over the course of File 686. These measures have a strong positive correlation (0.95) over all files.



Word_t/t_filter_10000 and Lexical.Density_value over the course of File 683. These measures have a weak positive correlation (0.1) over all files.



Revision_t/t_filter_10000 and Lexical.Sophistication.NAWL_value over the course of File 635. These measures have no correlation (0.0) over all files.

Lexical diversity is correlated with separation time, which suggests that a larger lexicon requires more planning time. Additionally, lexical diversity is correlated with both jumps and chunks, which implies that some revision time is dedicated to diversifying the lexicon.

Furthermore, jumps are correlated with syntactic complexity, while chunks are not. This indicates that more complex syntax requires more revision at the initial time of writing the sentence rather than after the fact.

Conclusion

Keystroke logging does offer a method to discover which writing strategies and cognitive processes produce the most complex text, as it is possible to connect characteristics of the writing process to lexical and syntactic complexity of the end-product text. The results suggest that in-sentence revisions (jumps), post-writing revisions (chunks), and longer pauses between words and sentences (separator_t/t_filter_10000) are all indicators of more complex text. More research should be done regarding the cognitive processes that are related specifically to revising and planning while writing, as well as how they relate to text complexity.