CREATING A FLUENCY-BASED MODEL CLASSIFYING TYPOGRAPHIC REVISIONS IN L1 AND L2 WRITTEN TASKS

FENNA BRONWASSER

STUDENT NUMBER: 2030650

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY

SUPERVISOR : ir. Rianne Conijn & dr. Drew Hendrickson

SECOND READER : dr. Martin Atzmüller

TILBURG UNIVERSITY SCHOOL OF HUMANITIES AND DIGITAL SCIENCES DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

TILBURG, THE NETHERLANDS

JANUARY 10th, 2020

Preface

I would like to acknowledge everyone who helped me in completing my thesis, which hopefully allows me to take a big step towards finishing my master degree in Data Science. Firstly, I would like to thank my friends and family for supporting and encouraging me. Secondly, my peers who were there for me when I needed help with the many glitches in my code; My mother with helping with the annotation; Sander v. Hulten for providing correction and feedback on my writing; Luuk v. Waes for coming to my proposal presentation, and offering very useful comments afterwards. The participants of the experiments for providing the valuable data used in this thesis; Drew Hendrickson for coordinating the thesis and allowing and helping me to switch thesis topic and supervisor at the last moment; And most notably, I would like to acknowledge and thank my thesis supervisor Rianne Conijn, who guided me through the process, giving great feedback along the way and being available when I needed assistance.

CREATING A FLUENCY-BASED MODEL CLASSIFYING TYPOGRAPHIC REVISIONS IN L1 AND L2 WRITTEN TASKS

Abstract

Typographic errors, although being minor incidence, can influence the writing process by breaking the linear writing flow. Classifying typographic revisions could function as a first step towards further analysis on reducing the undesired effects of typographic errors, or for the purpose of filtering typographic revision; as typographic errors are non-deliberate, mechanical errors which you might not want to include when analyzing other types of revisions. This study is a continuation of the research of Conijn, Zaanen, van Leijten, & Van Waes, 2019. Whereas previous studies on typographic/typing errors focused mainly on the finished writing product, this study utilizes a process-based approach which allows for the exploration of new features. In contrast to the research of Conijn et. al., (2019), this research focusses on typographic revision instead of typographic errors, trains the model on writing tasks which have a more natural setting, compares the typographic revision classification between first language writers and second language writers, and uses fluency-based features for building a classification model. Results show that the fluency-based model was reasonably effective in classifying typographic revision. A difference in performance of the model between first and second languages writers was found, however it is unclear whether the dissimilarity in language and language acquaintance accounts for this performance difference.

1. Introduction

A brief slip of the finger, and accidentally the wrong keyboard key is pressed; it happens to everyone who frequently writes on a keyboard. While typing, it is easy to forget that writing is a very complex cognitive process (Sweeney, 1997) in which many factors are involved including planning, transcribing and reviewing (Hayes & Flower, 1980; Ronald T Kellogg, 1996). A typographic error (or 'typo') is defined as an error due to mechanical failure or 'slip of the finger' (Stevenson, Schoonen, & de Glopper, 2006). Even though typographic errors appear to be minor incidents they still require cognitive processing to revise, and can disrupt the linear writing flow (Conijn, Zaanen, van Leijten, & Van Waes, 2019). Interruptions due to typographic error revise could divert the writer from the writing process and as a result potentially lower the quality of the writing product.

Writing in the present day is an incredibly important mean of conveying information. In academia, writing is fundamental for the development and generation of knowledge (Breuer, 2015), and it is hard to

imagine academics without written pieces. Since the quality of the writing product relies heavily on the writing process (Graham & Sandmel, 2011), it is key that the writing process is sound and that the writing process can happen reasonably uninterrupted (Leijten, van Waes, & Ransdell, 2010). Before further research on the effects of typographic errors on the writing process can be carried out, it is essential to be able to identify the typographic errors and being able to separate them from other types of errors, which is the aim for the current study.

For this study, keystroke data will be utilized. Keystroke data is gathered using key logging software, which records every keystroke made on a keyboard by a user. Key logging is a powerful tool for gathering data on multiple aspects of writing and makes detailed information on writing processes accessible (Miller, Lindgren, & Sullivan, 2008). Keystroke data (or keystroke/key logging) makes it possible to detect (typographic) errors when they are being made, in contrast to analyzing writing errors in the final product. Analyzing typographic errors in keystroke data gives insight on features such as timing and revision, which will be utilized as features to build the classification model.

The current study will continue on the research of Conijn et. al. (2019), in which they have built a model that was successful in identifying typographic errors in some circumstance. Yet, the model was only slightly better than the baseline in detecting typographic errors in source-based writing tasks. Similar to Conijn et. al. (2019) a gap was identified in the existing literature, as previous researched typically focused on identifying (typographic) writing errors in the end product in contrast to the writing process. Most studies researching the topic of (typographic) error focus on automatic correcting within the writing product (e.g. spellcheckers) (Whitelaw, Hutchinson, Chung, & Ellis, 2009), however these studies often do not distinct different types of errors nor do they try to link the error to the underlying cognitive processes behind the making, detecting and correcting of typographic errors. Therefor this study aims to improve upon the model of Conijn et. al. (2019) in identifying typographic errors within the writing process.This study expands upon the previous model (Conijn et. al., 2019) in four ways;

1. The previous model was trained on copy-task data in which every deviation from the original text could be flagged as an error, including non-revision errors, using a lexicon. The current study will try to improve upon performance by training on a source-based writing task. This change is made since the model that was trained on the copy-task dataset did not generalize well to the source-based writing task, indicating that the copy-task data might not be a sufficiently natural writing setting for the model to be trained on. Therefor the current model is trained on the source-based writing with is more comparable with academic writing
2. Whereas in the previous study, every disparity between the original text and the copy-task could be marked as a typographic error, there current study will solely look at revisions, and therefor attempt to identify typographic revisions (typographic errors that are revised) instead of typographic error.

Most often typographic errors are identified using a lexicon, where the misspelled word is compared to possible intended word. However according to Wengelin (2007), the majority of typographic errors are detected and corrected relatively fast. Presumably, when typographic errors would be analyzed in the writing product, most would already be corrected, indicating that most typographic errors, since they are revised, will be included in the analysis.

3. The current study will utilize two datasets, one of which are first language writers (L1) the other second language writers (L2). On average, L2 writers make more errors when writing than L1 writers (Doolan, 2017; Eckstein & Ferris, 2018). Although the data sets differ in language, are L1 vs L2 and have a slightly different task, the task and the dataset considered to be similar enough for comparison. Using both written tasks from L1 and L2 writers could provide insight into the differences in typographic errors between the two groups and furthermore, assesses the generalizability of the model.
4. Whereas the previous model used bigrams features and inter-keystroke interval (IKI) for prediction, the new model will use fluency features. Although there is no clear definition of writing fluency, within the scope of this study it is characterized by writing production and tempo. The cognitive process behind making an error and revising them varies between different types of errors (Larigauderie, Gaonach, & Lacroix, 1998), this study hypothesizes that because of the difference in the cognitive process the changes in fluency will also be affected differently. Analysis on fluency can shed a light on the underlying processes which take place during writing (Chenoweth & Hayes, 2001) and giving insight on writer's attention and cognitive processing (Chenoweth & Hayes, 2001; Schilperoord & Sanders, 1999).

Two models will be trained and tested, one will be trained on the L1 dataset and the other the L2 dataset, afterwards the L1 model will be tested on the L2 dataset and vice versa. This will provide insight in whether, and to what extent the models are independent of language and L1/L2 factors. The models will be assessed on their accuracy to correctly classify typographic revisions from other writing revision.

2. *Literature review*

Although this study is largely computational, the features of the model will be established based on the construct fluency and writing errors which need to be addressed in a theorical framework.

W*riting process*

Writing is very cognitively demanding; this has been explained in several models regarding the working memory in writing (Ronald T Kellogg, 1996; Shah & Miyake, 1996). While writing a text there are many cognitive processes taking place such as planning, monitoring, reviewing, retrieving, and transcribing (Abdel Latif, 2013; Hayes & Flower, 1980). Often, much more time is spent planning, monitoring, reviewing, and retrieving than the actual 'writing' (/transcription) (Uppstad & Solheim, 2007). However, transcribing is often seen as the main attribute of writing.

Kellogg (1996), has created a model explaining how the writing process takes place in the working memory. Within this model it is explained how different task take up separate spaces (so called slav-systems) in the working memory. To illustrate this, if there is a verbal concurrent task it affects the short-term storage of verbal information but not that of visuospatial information, and vice versa. Writing and the subprocesses behind it are very demanding on the working memory, and the capacity of the working memory is limited (Hayes & Chenoweth, 2007). Writing is a constant process of selective and divided attention, switching attention, task switching, coordination of the slav-system (Baddeley, 1996). The effects of these underlying processes of writing display themselves in frequency or duration of pauses, and changes in fluency, which are all reliable indicators of the cognitive management of processes (Fayol, 1999). In keystroke logging these indicators can be analyzed, therefor this type of data might give insight into the management of processes of a specific writer.

Just like others tasks within writing, detecting and revising an error need to be processed by the working memory, and therefor cause a pause and/or change in fluency (Larigauderie et al., 1998). Larigauderie et al. (1998) also found that the type of error (elaborated on further in the chapter) plays a role in the amount of memory resources required, indicating that, a typographic error could potentially surface differently within the keystroke data from other types of errors.

Fluency

Fluency is a term which is used frequently within linguistic but its definition and operationalization varies greatly between studies.

> "Written fluency is not easily explained, apparently, even when researchers rely on simple, traditional measures such as composing rate. Yet, when any of these researchers referred to the

> term fluency, they did so as though the term were already widely understood and not in need of any further explication" (Bruton & Kirby, 1987, p. 89).

In everyday conversations, fluency often refers to the ability to speak, read or write in a language which is foreign to them, however this differs from the meaning of fluency in academic literature. Although fluency is sometimes used as a broader construct, denoting a more general proficiency in a language, including defining it as the 'richness' of the writers (Abdel Latif, 2013). More often, fluency is defined as being "… the learner's capacity to produce language in real time without undue pausing or hesitation" (Skehan, 1996, p. 22), in which writing productivity and speed are the primary factors. The latter definition is employed in the current study.

Fluency in writing can be measured in both the writing product or the writing process, though analyzing the writing process can provides us with more (potential) indicators of fluency. There is no absolute consensus on which parameters need to be taken into regard in order to measure fluency reliably, since the factors utilized varied considerably between studies. Even though there is a lack of consensus, the general understanding is that fluency is measured by pausing amount and duration, the amount of revisions and the production rate (Chenoweth & Hayes, 2001; Van Waes & Leijten, 2015), with words-per-minute being a highly used measurement. Not only words-per-minute can be a measurement of speed, but also typing speed (time between two characters) can also be used as a parameter for fluency.

Another noteworthy indicator of fluency is writing bursts, a writing burst being defined as a period of active writing, followed by a pause (P-burst) or revision (R-burst) (Van Waes & Leijten, 2015; Chenoweth & Hayes, 2001;2006). In their study, Chenoweth and Hayes concluded that burst length is "fundamentally related to fluency" (p.94), in which burst length is positively linked to fluency. Within the study of Van Waes and Leijten (2015) they explored a large number of indicators of fluency and their validity. They found four main components of fluency which were considered effective indicators of fluency, namely; (1) production, (2) process variance, (3) revision, and (4) pausing behavior. These components, except process variance, will be operatized within the current study, the way these are analyzed (e.g. production being analyzed as words-per-minute or character-per-minute) will be discussed in the method section. A brief overview of variation in operationalization of fluency between studies is provided in the table 1.1, note that this list is not exhaustive.

Table 1 – brief, non-exhaustive overview of the features that are used to measure fluency

Article	Feature(s)
(Chenoweth & Hayes, 2001)	o Composition rate (words p minute) o P-burst length o R-burst length
(Fellner & Apple, 2006)	o Composition rate (words p minute)
(Kellogg, 1987)	o Composition rate (words p minute/total)
(Bruton & Kirby, 1987)	o Composition rate (words p minute)
(Ballator, Farnum, & Kaplan, 1999)	o Holistic scoring (non-quantitative)
(Rosenthal, 2006)	o Number of correctly spelled words written o Number of sentences written o Number of letter sequences

(Van Waes & Leijten, 2015)	o Production: mean number of characters (incl. spaces)
	  o during the process (per minute)
	  o in the final product (per minute)
	  o per .10 interval (corrected for abs. opt. of 400 cpm)
	  o per P-burst (per minute; threshold 2000 ms)
	o Process variance: st. dev. of characters (incl. spaces)
	  o per .10 interval (per minute)
	  o per .10 interval (corrected for task maximum)
	o Revision: mean number of characters (incl. spaces)
	  o product vs. process ratio
	  o length of R-burst
	o Pausing behavior
	  o mean pause time length between words (th > 200 ms)
	  o proportion of total pause time (th > 2000 ms)

(Matsuno, Sakaue, Morita, Murao, & Sugiura, 2007)	o Pause time between words
(Kaufer, Hayes, & Flower, 1986)	o Pause time between words
(Johnson, Mercado, & Acevedo, 2012)	o Total number of words o Average sentence length
(Baba & Nitta, 2014)	o Number of words per 10 minutes
(Olive, Alves, & Castro, 2009)	o P-burst (length and frequency) o R-burst (length and frequency)

(Typographic) errors

There are various kinds of errors that can be made during writing, including spelling, typographic, grammatical, semantic, lexical, and stylistic ones (Islam & Inkpen, 2011). However, typographic errors are slightly different from other writing errors, in that they are accidental, non-consciously made errors, a 'slip of the finger' (Wengelin, 2007). The cognitive processes behind typographic errors (e.g. compared to semantic and syntactical errors) seem to differ, as typographic errors are put less strain on the working memory (Larigauderie et al., 1998; Wengelin, 2007).

Similar to the study of Conijn et al. (2019), the definition of typographic errors proposed by Stevenson, Schoonen, & de Glopper (2006) will be utilized. Stevenson, Schoonen, & de Glopper (2006, pp. 230–231) recognized five possible cases of errors which could be considered typographic errors;

a) The pre-revision form does not conform to the orthographic rules of the language (e.g. ''mooore'' instead of ''more'').
b) The pre-revision form involves a letter string which does not conform to a likely pronunciation of the word (e.g., ''improant''instead of ''important'').
c) The semantic context indicates that the pre-revision form could not have been intended (e.g., 'I got apresent form my mother'' instead of ''I got a present from my mother'').
d) The same word is written correctly at an earlier point in the text.

In conclusion, the fluency of the writing can reveal a lot about the various writing processes interacting and how diverse tasks place a strain on the working memory of the writer (Larigauderie et al., 1998; Wing & Baddeley, 2009). The typographic error differs in cognitive processing from other revisions (Larigauderie et al., 1998), therefor fluency could potentially be a suitable parameter to use in a model for classifying typographic errors within writing revisions.

L1 and L2 writers

Even though an L2 speaker can master a languages very well, a majority of the cases, a fluency gap is found between L1 and L2 writers/speakers (Breuer, 2015; Van Waes & Leijten, 2015). Studies have showed that fluency, specially bursts, show a clear distinction between L1 and L2 writers since the it is more cognitive demanding to both translate and handle (complex) language structures (Chenoweth & Hayes, 2001) . "The length of pause—and revision—bursts drops significantly when (young) writers compose text in a second language (L2), as opposed to text production in their mother tongue." (Van Waes & Leijten, 2015, p. 81).

Based on the literature it is expected that there will be a difference between the L1 and L2 writing tasks in errors and fluency, however there no clear understanding of whether L2 make and revise a typographic error differently from L1. In theory are typographical errors separated from language skills, skills

which form the gap between L1 and L2 writers. Therefor it is supposed that there will be minimal difference in typographic errors. However, other unknown factors could potentially play a role which would result in a disparity between L1 and L2 writers in making typographic errors.

3. Research Questions

This study aims to construct a model intended for recognizing typographic revisions during the writing process using fluency. The following research questions were composed:

1) Can a model accurately classify typographic writing revisions from other writing revisions using fluency features?
2) Does the effect of typographic revisions on fluency differ substantially between L1 and L2?

4. Methods

To answer the research question, an analysis is performed using two different keylog datasets, one being a source-based writing task, written in the native languages of the participants (Dutch), the other one being an academic summary task, written in the second languages of the participant (English). The revisions needed to be annotated manually, after which the two data sets were cleaned and transformed for the feature extraction, subsequently, 17 fluency features were attracted for the classification of the revisions. After the participant variance was accounted for, four different classification algorithms were trained on the two datasets separately, totaling 8 models. All models were tested on both datasets to evaluate the generalizability of the models.

Source-based writing task dataset (L1)

The first dataset (L1) is the source-based writing task written in Dutch, which was collected and utilized within the research of Leijten et al. (Leijten, Van Horenbeeck, & Van Waes, 2019; Leijten, Van Waes, Schrijver, Bernolet, & Vangehuchten 2019). Participants (N = 49) where Dutch graduate students, between 21 and 48 years of age (M = 27.4, SD = 8.1), and mostly (73%) female. These participants were requested to write a text of 200 to 250 words on humanitarian aid, renewable energy, climate change, or animal rights. Based on the topic that was chosen sources where provided, the three sources being: report, a web text, and a newspaper article. Participants were given 40 minutes to finish the task and where able to consult the internet. In total, 66 source-based writing tasks were collected with a mean length of 2,980 characters (SD = 1,205) and made 116 revisions (SD = 62). Keystroke data were collected using Inputlog software.

Academic summary task dataset (L2)

The second dataset (L2) is an academic summary task written in English, in which students (N = 90) were asked to summarize a journal article in 100-200 words within an experimental setting. Students were enrolled in an English course for second language learners, and provided informed consent on partaking in the experiment. The journal article was an article by Woong Yun & Park (2011), which was a 2 × 2 experimental design study within the field of Communication and Information Sciences. Student got 30 minutes to write the summary after reading it, a similar timeframe compared to the source-based writing task. An average of 95 (SD = 40) revisions were made, and the average length of the final task was 1325 characters (SD = 447). The keystroke data were collected using Inputlog software.

Annotation

Both data sets needed to be annotated manually. A revision was only mark if the backspace key was used, deletions or replacements were not included. During the annotation the revisions needed to be marked as either a typographic revision or non-typographic revision (including both non-typographic errors, such as spelling or grammar errors, and deep revision). The decision on whether a revision was marked as typographic or non-typographic was made based on a comparison between original text and its replacement. To distinguish typographic errors from other writing errors, the coding guide of Stevenson et al. (2006) was used combined with additional annotation rules specially for the datasets that are utilized. The coding guide can be found in Appendix A.

To test the reliability of the annotation a second annotator was taken on, who was instructed and provided the coding guide. A small subset of 729 revisions was annotated by both annotators independently. To assess the reliability the inter-rater reliability was calculate. An inter-rater reliability of 88.6% was computed. Points of discussion were, for example, when the first letter of a word was revised or when it was unclear whether a participant knew how to correctly write the word. When there was a serious doubt about the nature of the revision, the revision was marked as non-typographic. During the annotation phase it was uncovered that two participants of the L2 dataset did know form coherent words and sentences, indicating that these were non-serious attempts, the two participants were removed

Data Cleaning

After annotation, the datasets were loaded into R and cleaned in preparation for the analysis, the following steps were taken; (1) within the L2 dataset there was an additional task included, this was a copy-task, which could not be used was removed from the data set manually; (2) within the L2 dataset the mouse movements were recorded and included, these were removed from the dataset. This did however create a

gap and a pause which needed to be corrected afterwards by recalculating certain temporal variable; (3) participants were allowed to copy and paste during the experiment, and because of the nature of the task, a lot of participants did copy and paste a significant amount of text, which skewed the data by having long pauses and not producing many characters, a total of 5 participants were removed because of this (one from the L1 dataset and 4 from the L2 dataset);(4) In the data from one participant (L1) an error was found (impossible time values), this could not be correct, leading to removal; (5) Before the source-based writing task (L1), the participants had also completed a copy-task, this task was already removed from the dataset, however this did cause a gap in time related variables which needed to be corrected; (6) some revisions were left unmarked or marked incorrectly (e.g. 'nn' instead of 'n') during annotation, these were corrected.

After cleaning 173683 rows of the original 186007 of the L1 dataset were left, and 208276 rows of the original 401157 of the L2 dataset were left. Quartiles and boxplots were used to look at possible outliers, although a high variability was found, no further action was taken to remove outliers.

Feature extraction

From both datasets the same fluency features were extracted, which were chosen based on the findings in the literature. The annotated revisions are extracted from the data including 20 keys prior and 20 keys post the revision. This 20-key limit pre and post is chosen since if it was shorter the r-burst, p-burst and number of pauses variables would not provide any useful information, as there would be few to none pauses and burst within a 10 or 5 key window. The key-window also should be not too long as the effects of the revision on fluency decline. Table 2 provides an overview of the 19 features that were extracted and how these are operatized, including whether the feature was extracted from the 20 pre', 20 post-revision keys, during the revision itself or overall (overall meaning 20 pre- 20 post-revision keys, plus the revision itself). To better understand the different time feature within keystroke data an overview is provided in an image of Conijn et. al. (2019). Table 3 lists the means and standard deviation of all features (except factors), note that the pause time between keys drastically higher is in the L2 dataset, this is caused by outliers, when a trimmed mean or median is used they are evenly balanced (median pause time between keys L1 = 138 & L2 = 144).

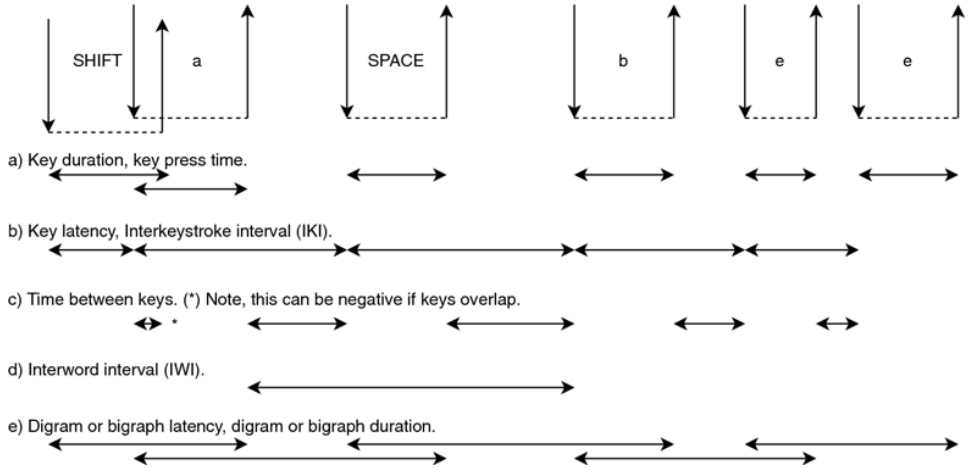

Fig. 2 Time-based features extracted from the keystroke log of typing: "A bee" from Conijn, R., Roeser, J. & van Zaanen, M. Read Writ (2019) 32: 2353.

Next to fluency features, character bigrams were also extracted, character bigrams (from now on referred to as bigrams) are combinations of two characters within a string of tokens/characters. Bigram features were a large component of the classification model within the previous research of Conijn et. al. (2019). For the purpose of this research they were extracted to evaluate whether a bigram feature is beneficial to the current models, and to investigate whether potentially expending the model with additional bigram features would improve model accuracy.

The way in which the bigrams are utilized in this study is not fully consistent with how they were utilize in the previous research. Since their model was trained on copy task data, they were able to compare the extract the expected bigram (the bigram that it should have been), the typed bigram and the swapped bigram. For each of the four bigrams they extracted four features: bigram frequency, repetitiveness, adjacency, and hand combinations. For the purpose of the current study, only the frequency of typed bigrams was extracted and included in the models.

The data on the frequency of the bigrams were provided, both in English and Dutch, through the previous research. The frequency of the bigrams was calculated alike Van Waes, Leijten, Mariën, & Engelborghs (2017). The frequency was divided in three categories, the 30% most, frequent bigrams were classified as high-frequent, the bottom 50% frequent bigrams was classified as low-frequent, the remaining bigram were marked as medium-frequent.

Table 2 – which features are extracted and how it is operationalized

Feature	Operationalization	Overall	Pre-re-vision	Intra-revision	*Post-re-vision*
Average pause time between keys **5 keys (pre – post)** **20 keys (pre – post)**	See key latency (fig 2), the time between the start of one key press until the next start key press. The pause time was summed and divided by the number of keys to retrieve the average pause time.	x	x	x	x
Production rate per 10 sec	For every 10s the amount of characters produced (incl. spaces and interpunction, not incl. shift, ctrl and copy paste) was calculated.	x	x		x
Length of the revi-sion	Number of the backspace was pressed (how many characters are revised)			x	
Number of pauses	A pause time was calculated based on the mean key latency (see fig 2) of a participant plus two standard deviations		x		x
Length of R-burst	The amount keys pressed between a pause (see nr of pauses for definition) or a revision and followed by a revision.	x			
Length of P-burst	The amount keys pressed between a pause (see nr of pauses for definition) or a revision and followed by a pause.	x			
Interword interval	See fig 2., time between the end time of last key of a word and the start of the first key of the following word. The inter-word interval was only extracted when the space key was pressed	x			
Other revisions	The sum of (other) backspace keys pressed within 20 keys pre and post revision	x			
Bigram frequency	The bigram was extracted by using the last two character be-fore a revision, or if the last character was equal to the first character after the revision (because too much was revised), the 3rd and 2nd characters before the revision were used. The bigram was linked to the frequency table and was categorized as high frequency, low frequency and medium frequency, us-ing dummy coding. It there was a punctuation or modifier key (e.g. ctrl) the bigram frequency was not marked.		x		

Table 3 – feature mean plus sd **(**pre = pre revision, post = post revision)

Feature	L1 - Mean	L2 - Mean
Production rate per 10 sec pre	34.93 *(sd = 34.19)*	32.15 *(sd = 31.51)*
Production rate per 10 sec post	35.31 *(sd = 32.36)*	32.05 *(sd = 30.87)*
Production rate per 10 sec overall	24.51 *(sd = 23.40)*	22.79 *(sd = 21.44)*
Length of the revision	16.23 *(sd = 18.01)*	14.88 *(sd = 16.25)*
Number of pauses pre	3.23 *(sd = 2.08)*	2.79 *(sd = 1.88)*
Number of pauses post	3.23 *(sd = 2.19)*	2.77 *(sd = 1.91)*
Average pause time between keys 5 keys pre	285.42 *(sd = 362.20)*	703.08 *(sd = 2333.43)*
Average pause time between keys 5 keys post	303.69 *(sd = 404.51)*	877.90 *(sd = 2902.69)*
Average pause time between keys 20 keys pre	277.41 *(sd = 212.30)*	727.03 *(sd =1404.285)*
Average pause time between keys 20 keys post	294.90 *(sd = 214.35)*	792.99 *(sd = 1476.90)*
Average pause time between keys inter revision	490.87 *(sd = 857.18)*	1402.787 *(sd = 6022.75)*
Average pause time between keys overall	305.10 *(sd = 169.05)*	815.62 *(sd = 1084.60)*
Length of R-burst	4.23 *(sd = 4.34)*	4.24 *(sd = 5.08)*
Length of P-burst	5.36 *(sd = 4.64)*	6.18 *(sd = 5.41)*
Interword interval	244.77 *(sd = 330.84)*	736.37 *(sd = 3089.82)*
Other revisions	6.37 *(sd = 4.61)*	6.71 (sd = 4.50)

Analysis - Computing Participant averages

If the absolute features would be utilized in the models, the models would be influenced by the typing characteristic of each participant (e.g. revisions of a participant who is a slow writer could be flagged as non-typographic revision because of his writing style instead of revision characteristic). There is a lot of variation between participants in how they write, for example the lowest average pause time between keys of a participant (L2) was 100ms, and the highest 323ms. To account for this variance the average of the participant was calculated and subtracted from the features.

The final data set for training and testing the models consist of a row for each revision and columns for each feature (as well as participant ID and output variable ‘typo.’). A table of participant averages is

looped through this data set, leaving only the deviation from the participant's average. As seen in Table 3 the average pause time between keys is much higher in L2 dataset than the L1 dataset; This is however due to due outliers, meaning that if you would subtract the mean from the feature, it would result in almost all values being negative. To correct for this a trimmed mean (0.05) is used for as more robust metric. The pause threshold ('how long before you can call it a pause?') is also dependent on the participant since it uses mean + (2 * standard deviation) as the threshold. Because the ranges differ (greatly) between the features the numeric features are normalized. A total of 5506 errors are retrieved from the L1 data set and 8093 from the L2 data set.

Table 4 – average pause threshold and time between keys

	Pause threshold	**Time between keys**
L1	415.83 *(sd = 87.24)*	181.01 *(sd = 36.10)*
L2	518.36 *(sd = 196.05)*	173.52 *(sd = 37.06)*

Analysis - Classification model

The goal of the automatic model is that the revisions need to be successfully classified into typographic or non-typographic. To do this firstly both the L1 and L2 data set were split into train and test (80% / 20%) using the 'createDataPartition' function from the caret package (Kuhn et al., 2018) which automatically tries to balance the distribution within the split of the outcome factor. Within the L1 trainings and test set is 40.6% of the total revision are typographic, for the L2 data set this is 52.5%.

For the training of the models, a ten-fold cross validation is utilized, using the 'trainControl' function from caret (Kuhn et al., 2018). During the decision of which machine learning models to use there we two things that needed to be considered; (1) the correlation between the features, within the L1 dataset there are 16 instances of a feature correlation >0.5, particularly between the production and pause between key features. This was expected and since the features all have a temporal element, it is hard to avoid. The correlation does however mean that for instance Naïve Bayes is not a good fit for the data; (2) The limited data point (in the L1 data set there is less than 3000 typographic revisions), which might be too small for e.g. a neural network. Ultimately, KNN, Random Forest, Logistic Regression and Support Vector Machine were chosen. KNN and Logistic Regression are overall simpler and made it fairly easy to uncover the effect of each variable, however Random Forest and Support Vector Machine are more likely to work because the data is not vey clean, with some features having an extremely high variance and overall not being very balanced.

The ‘train’ function in the caret package is applied for the model training, using the methods ‘knn’, ‘rf’, ‘glmnet’, and ‘SVMlinear’. The metric is set to ‘Accuracy’ in all cases. The ‘family’ in glmnet is set to ‘binomial’, for SVM, ‘preProcess’ is set to "center" and "scale" with a tune length of 10. All other are set to default. Afterward there is a total of 8 models trained, 4 on each data test. The models are then tested on the both the L1 and L2 test sets, using the ‘predict’ function from caret. were trained on the F-score, and run using 10-fold cross-validation. As evaluation metrics, precision, recall, and F-score are reported, this allows for the interpretation of whether the models have more issues with false negative or false positives.

Lastly the best model is repeated but the bigram features are removed to see whether they contributed to the model and if expending on bigram feature will result in a better performance. All other metric and option and consistent with the previous description.

5. Results

In all instances the Random Forest performed best, when the model was tested on its own source (e.g. L1 model predicting the L1 test set) the recall and precision were fairly balanced. However, when tested on the other language it the recall when up, meaning more typographic revision were flagged, but the precision went down, indicating that a lot of non-typographic revisions were flagged as typographic revisions. Dependent on the objective of the classification is a high recall or a high precision might be preferred. If the goal is to remove typographic revision, a high recall might be favored as this means that a high percentage of typographic revision is removed although if the precision is low many non-typographic revisions will also be removed.

The majority class baseline for L1 is 0.594 and for L2 this is 0.522, which was improved upon. Note that the upper baseline is 0.886 based on the interrater reliability. Were as the results from the L1 model – L1 test are fairly stable between the model types, the L1 model – L2 test results varied greatly between model types where Random Forest picked up almost all true positives, the recall from KNN was only slightly higher than the majority class baseline.

The model without bigram features performed slightly less than the model with bigram features. While only one feature of one bigram is now included in the model, there is a (slight) performance difference.

Table 5 – Overview of the L1 model with and without bigram frequency features tested on both L1 and L2

	L1 model – L1 test			**L1 model without bigrams – L1 test**		
Model	Recall	Precision	F1-Score	Recall	Precision	F1-Score
Random Forest	**0.7979**	**0.7811**	**0.7979**	**0.7917**	**0.7845**	**0.7881**
Support Vector Machine	0.7657	0.7519	0.7587			
Logistic Regression	0.7688	0.7404	0.7543			
KNN	0.7232	0.7033	0.7232			
	L1 model – L2 test			**L1 model without bigram – L2 test**		
Model	Recall	Precision	F1-Score	Recall	Precision	F1-Score
Random Forest	**0.9618**	**0.4919**	**0.6509**	**0.8790**	**0.4982**	**0.6359**
Logistic Regression	0.6115	0.6621	0.6358			
KNN	0.6815	0.5679	0.6196			
Support Vector Machine	0.5669	0.6784	0.6176			

Surprisingly, the L2 model performed better on the L1 test set, however it needs to be taking into consideration that the distribution on the revision was not equal. Similar to the L1 model – L2 test results, the L2 model – L1 test results report a large recall and low precision.

Table 6 – Overview of the L2 model with bigram frequency features tested on both L1 and L2

	L2 model – L2 test			**L2 model– L1 test**		
Model	Recall	Precision	F1-Score	Recall	Precision	F1-Score
Random Forest	**0.7172**	**0.7744**	**0.7447**	**0.9587**	**0.6210**	**0.7538**
Support Vector Machine	0.6102	0.7247	0.6625	0.9831	0.6033	0.7478
Logistic Regression	0.6446	0.6829	0.6632	0.9877	0.6011	0.7473
KNN	0.5745	0.7467	0.6494	0.9587	0.6210	0.7538

To evaluate the features, the feature importance of the best model (the Random Forest model trained on the L1 data set) is computed. It is evident that within the model, mainly the pause time between keys contribute a great deal to the model, except for the pause time 20-key pre and after revision. Which could be an indication that after 20-keys the effect of the revision on fluency fades. Although calculating the average P-burst and R-burst length over a small snippet is not ideal, as is visible through the large frequency of empty cells, still both the R-burst and P-burst contribute considerably to the model.

It is clear that the bigram features contribute the least towards the model, however it must be considered that the features that were used were a fraction of the bigram features used within the previous study.

Table 7 – Feature importance of Random Forest model trained on the L1 set

Feature	L1 – Random Forest
Average pause time between keys 5 keys post	100.000
Average pause time between keys inter revision	62.282
Average pause time between keys overall	37.492
Average pause time between keys 5 keys pre	36.852
Length of R-burst	36.341
Length of P-burst	31.029
Production rate per 10 sec overall	29.183
Interword interval	26.895
Average pause time between keys 20 keys post	25.169
Production rate per 10 sec post	24.983
Average pause time between keys 20 keys pre	22.923
Length of the revision	21.970
Production rate per 10 sec pre	21.724
Other revisions	14.994
Number of pauses pre	10.724
Number of pauses post	9.233
High frequency bigrams	5.039
Low frequency bigrams	1.945
Medium frequency bigrams	0.000

6. Discussion

The aim of this study was to create a classification model that accurately classifies typographic error revisions versus non-typographic revisions. Although the model was a successful continuation of the study of Conijn et al., (2010), and supports that fluency-based model can classify typographic revision, there were some validity and reliability issues that needs to be addressed.

Even though the fluency features performed well in the model, is must be considered that the features were very noisy, with many extreme outliers and a high variability. In the literature review the link between cognitive processing and fluency was addressed, and the model partially supports this notion, though error revision processing is not the only cognitive process that is measured when analyzing fluency, participants could also simply be distracted, or reading another line within the text. The fact that in both tasks the participants had to constantly switch between reading and writing also had a large impact on the fluency features making the data not entirely suitable for training a model for typographic revision classification. Leaving out scrolling and cursor movements created big gaps in the data consequently causing extreme outliers. Leaving in the mouse movement also is not a suitable option since the pause times between mouse movement are often close to 0ms, this does not correspond with the normal pause time between key presses.

Another problem with the nature of the task was the amount of copying and pasting, although this is true to actual (academic) writing, some participants copied more than they wrote, going back and forth between copying and revision instead of a linear writing flow. What also happened frequently is that multiple revisions were followed by each other. If these are removed a lot of data is lost, but it is difficult to measure the influence of revision on fluency if the fluency is not yet recovered from the previous revision.

In the literature it was found that r-bursts and p-burst are good indications of fluency, as it is believed that fluent writing require less pauses in their writing to think or revise. Unfortunately, this a feature that is hard to operatize, since one or two p- or r-burst after (or before) the revision, the effect of the revision on fluency has probably faded. The problem with this is that one or two r- and/or p-bursts are not a very constant, reliable variable to use for the model.

In addition to the predictor variables being noisy, the output variable was also less reliable than preferred, even though an interrater reliability of 88,6% is not unsatisfactory, still 12% of the revision might be incorrectly trained and tested. In retrospect, removing or not marking the revision that cause serious doubts might be a better option, at least for the trainings data in order to make sure the model(s) are accurately tained.

Although a large potion of participant variability was probably accounted for by removing the means, the variability of how a participant tackles a typographic error was not accounted for. While on average the cognitive process behind the revision followed directly after the revision, as seen in the data, in some

instance's participant process the error before revising (before backspace is pressed) or even during the revision (during the backspace is pressed). These variations between participants cannot be accounted for easily.

The problem of correlation also needs to be addressed as some features, expectedly, were highly corrected. Some of the highly correlated feature could most likely be removed, for future studies who would want to further investigate fluency features, a PCA is a suggested as this could be effective for feature reduction.

One of the biggest limitations of using solely fluency measures is that if a temporal noise occurs almost all features will be affected and the model has no other type of feature to then depend on. On that regard, the previous model that combined IKI and bigrams had an advantage. For future studies combining fluency with another type of predictor is recommend, for instance combining fluency and eye tracking, because an eye tracker can measure what the participant is reading and/or looking at, indicating what the participant is processing. The one bigram feature that was included did not perform adequately; still additional bigram features could be added to see what happens when there is more interaction between the bigram feature.

Although at first glace the data sets seem large, if considered what is left when extracting the errors/revisions is it not a great number of data points. Keystroke logging by nature is quite noisy data and one way to address this is by making sure the is enough data to adequately train the model. For follow up research, collecting more data is suggested.

7. Conclusion

The current study made great process in continuing the investigation of typographic errors/revision classification. Both models outperformed the majority class baseline and was not far removed from the upper limit of 88.6 (based on the Interrater Reliability). The findings in the literature review suggested that typographic error require a different cognitive process than other errors or revisions, and the current results support this notion. This research cannot isolate the cause of the difference in the effect of typographic and non-typographic revisions on fluency, however findings suggested that there is a link between fluency and revision type, confirming that it possible to build a classification model using fluency features.

It is not clear why the L1 underperformed on the L2 test set and vice versa. Although they still outperformed the majority class baseline the precision metric was remarkably low. This illustrates that the models are not generalizable to the other dataset, but multiple various factors could have led to the underperformance of the models. Such as the fact that the data was unbalanced, the difference is task between the L1 and L2 data set or the numerous outliers included in L2. The model cannot verify that the underperformance of the models when cross referenced is caused by the language or language acquaintance factor.

The aim of the study was to function as a steppingstone for typographic error/revision classification. Although the models created in this research are not accurate and generalizable enough to use in practice at the current stage, this research does provide new insights into what features might be effective for future classification models.

References

Abdel Latif, M. M. M. (2013). What do we mean by writing fluency and how can it be validly measured? *Applied Linguistics*. Volume 34, Issue 1, February 2013, Pages 99–105 https://doi.org/10.1093/applin/ams073

Baba, K., & Nitta, R. (2014). Phase transitions in development of writing fluency from a complex dynamic systems perspective. *Language Learning*. 64(1), 1–35. https://doi.org/10.1111/lang.12033

Baddeley, A. (1996). The fractionation of working memory. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.93.24.13468

Ballator, N., Farnum, M., & Kaplan, B. (. (1999). Trends in writing: Fluency and writing conventions. Holistic and mechanics scores in 1984 and 1996. *Princeton NJ: Educational Testing Service.*

Breuer, E. O. (2015). *First language versus foreign language: Fluency, errors and revision processes in foreign language academic writing*. *First Language versus Foreign Language: Fluency, Errors and Revision Processes in Foreign Language Academic Writing*. https://doi.org/10.3726/978-3-653-04262-7

Bruton, D. L., & Kirby, D. R. (1987). Research in the Classroom: Written Fluency: Didn't We Do That Last Year? *The English Journal*. https://doi.org/10.2307/818661

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in Writing: Generating Text in L1 and L2. *Written Communication*. 18(1), 80-98. https://doi.org/10.1177/0741088301018001004

Conijn, R., Zaanen, M., van, Leijten, M., & Van Waes, L. (2019). How to Typo? Building a Process-Based Model of Typographic Error Revisions. *Journal of Writing Analytics, 3*.

Doolan, S. M. (2017). Comparing patterns of error in generation 1.5, L1, and L2 FYC writing. *Journal of Second Language Writing*. https://doi.org/10.1016/j.jslw.2016.11.002

Eckstein, G., & Ferris, D. (2018). Comparing L1 and L2 Texts and Writers in First-Year Composition. *TESOL Quarterly*. Volume 52, issue 1, March 2018, Pages 137-162 https://doi.org/10.1002/tesq.376

Fayol, M. (1999). *From on-line management problems to strategies in written production.* Amsterdam: Amsterdam University Press.

Fellner, T., & Apple, M. (2006). Developing writing fluency and lexical complexity with blogs. *The Jalt Call Journal*.

Graham, S., & Sandmel, K. (2011). Graham 2011 Process Writing Approach Meta Analysis.pdf. *The Journal of Educational Research*, 104(6), 396-407. https://doi.org/10.1080/00220671.2010.488703 https://doi.org/10.1080/0022671.2010.488703

Hayes, J. R., & Chenoweth, N. A. (2007). Working memory in an editing task. *Written Communication.,* 24(4), 283–294. https://doi.org/10.1177/0741088307304826 https://doi.org/10.1177/0741088307304826

Hayes, J. R., & Flower, L. S. (1980). Identifying the organisation of the writing process. *Cognitive Processes in Writing*.

Islam, A., & Inkpen, D. (2011). An Unsupervised Approach to Detecting and Correcting Errors in Short Texts. *Computational Linguistics*.

Johnson, M. D., Mercado, L., & Acevedo, A. (2012). The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *Journal of Second Language Writing.* https://doi.org/10.1016/j.jslw.2012.05.011

Kaufer, D. S., Hayes, J. R., & Flower, L. (1986). Composing Written Sentences. *Research in the Teaching of English*.

Kellogg, R.T. (1987). Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory and Cognition*, (15), 256–266.

Kellogg, Ronald T. (1996). A Model of working memory in writing. In *The science of writing. Theories, methods, individual differences and applications*.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Huntt, T. (2018). caret: Classification and Regression Training. R package version 6.0-84. *R Package Version 6.0-79*.

Larigauderie, P., Gaonac'h, D., & Lacroix, N. (1998). Working Memory and Error Detection in Texts: What Are the Roles of the Central Executive and the Phonological Loop? *Applied Cognitive Psychology*. https://doi.org/10.1002/(SICI)1099-0720(199810)12:5<505::AID-ACP536>3.0.CO;2-D

Leijten, M., van Waes, L., & Ransdell, S. (2010). Correcting text production errors: Isolating the effects of writing mode from error span, input mode, and lexicality. *Written Communication*. https://doi.org/10.1177/0741088309359139

Matsuno, K., Sakaue, T., Morita, M., Murao, R., & Sugiura, M. (2007). Processing loads and fluency in writing: Comparison of the production fluency between native speakers and non-native speakers in terms of the 'cost criteria.' *Writing: Second Language*, (Nagoya Gakuin University, Nagoya, Japan).

Miller, K. S., Lindgren, E., & Sullivan, K. P. H. (2008). The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly*. https://doi.org/10.1002/j.1545-7249.2008.tb00140.x

Olive, T., Alves, R. A., & Castro, S. L. (2009). Cognitive processes in writing during pause and execution periods. *European Journal of Cognitive Psychology*, *21*(5), 758–785.

Rosenthal, B. (2006). Improving elementary-age children's writing fluency: A comparison of improvement based on performance feedback frequency. *Psychology - Dissertations.*, *24*.

Schilperoord, J., & Sanders, T. (1999). How hierarchical text structure affects retrieval processes: Implications of pause and text analysis. In *Knowing What to Write. Conceptual Processes in Text Production*.

Shah, P., & Miyake, A. (1996). The Separability of Working Memory Resources for Spatial Thinking and Language Processing: An Individual Differences Approach. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/0096-3445.125.1.4

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*, 38–62.

Stevenson, M., Schoonen, R., & de Glopper, K. (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing*. https://doi.org/10.1016/j.jslw.2006.06.002

Sweeney, L. T. (1997). The psychology of writing. *Acta Psychologica*. https://doi.org/10.1016/s0001-6918(97)00011-5

Uppstad, P. H., & Solheim, O. J. (2007). Aspects of fluency in writing. *Journal of Psycholinguistic Research*. https://doi.org/10.1007/s10936-006-9034-7

Van Waes, L., & Leijten, M. (2015). Fluency in Writing: A Multidimensional Perspective on Writing Fluency Applied to L1 and L2. *Computers and Composition*.

https://doi.org/10.1016/j.compcom.2015.09.012

Wengelin, Å. (2007). The word-level focus in text production by adults with reading and writing difficulties. In *Studies in Writing*.

Wing, A. M., & Baddeley, A. D. (2009). Righting errors in writing errors: The Wing and Baddeley (1980) spelling error corpus revisited. *Cognitive Neuropsychology*. https://doi.org/10.1080/02643290902823612

Appendix A

Stevenson, Schoonen, & de Glopper (2006, pp. 230–231)

a) The pre-revision form does not conform to the orthographic rules of the language (e.g. ‘‘mooore’’ instead of ‘‘more’’).
b) The pre-revision form involves a letter string which does not conform to a likely pronunciation of the word (e.g., ‘‘improant’’instead of ‘‘important’’).
c) The semantic context indicates that the pre-revision form could not have been intended (e.g., ‘I got apresent form my mother’’ instead of ‘‘I got a present from my mother’’).
d) The same word is written correctly at an earlier point in the text.

Additional coding rules for annotation of the L1 source-based writing task and L2 summary task:

a) If the revision is only the first letter of a word, you look at the proximity to the letter to its replacement character. If it is in a close proximity it is considered a typo, else it is a non-typo. (‘s’ instead of ‘d’)
b) If the first letter of a word is revised but is it equal to the second replacement character, it is considered to be a typo, else it is a non-typo. (‘a’ replaced by ‘path’)
c) If the revision was the Dutch (partial) word revised as an English word it is a non-typo (“of” replaced by “or”)
d) If a punctuation is revised it is only a typo when an obvious wrong key is used (“was he there/” instead of “was he there?”. If a point is revised as a comma (or vice versa) it is a non-typo
e) Unintentional capitalization or non capitalization is an typo (‘CAnada‘ revised by “Canade”)
f) Text revision that participants are not familiar with such as names are a non-typo
g) If it is not clear what the participant is revision it is a non-typo
h) if there is a clear indication (e.g. multiple revisions) that a participant has trouble spelling a ‘difficult’ (long/not much used) word it is a non-typo.
i) If there is a serious doubt the revision is marked as a non-typo.