Research Article

How to Typo? Building a Process-Based Model of Typographic Error Revisions

Rianne Conijn, *Tilburg University, The Netherlands and University of Antwerp, Belgium*

Menno van Zaanen, *Tilburg University, The Netherlands*

Mariëlle Leijten, *University of Antwerp, Belgium*

Luuk Van Waes, *University of Antwerp, Belgium*

Structured Abstract

- **Background**: The analysis of writing is complex, with planning, translating, and reviewing processes interacting in a non-linear fashion. Intuitively negligible activities such as the revisions of typing errors can have a large influence on the writing process, and hence also on the analysis of writing processes. For the analysis of writing, the importance of these errors and their revisions is twofold. On the one hand, they are low-level, and hence less-important types of revision, which we would like to filter, e.g., for more nuanced analyses of revision. On the other hand, typing errors, and especially the revision of typing errors, can (unwillingly) break the (linear) flow in writing. Therefore, it is important to identify these revisions to be able to determine their effect on disfluency and activation of other subprocesses.
- **Literature Review**: Previous work on typing errors commonly focuses on the writing product rather than the writing process. In this way, revised typing errors are omitted. In addition, no distinction is being made between typographic errors (slips of the fingers) and other types of errors. Some studies used keystroke data to manually annotate the revisions of typographic

errors during the writing task. However, the automatic analysis of typographic error revisions using process data is still understudied.

- **Research Questions**: Therefore, we aim to build a process-based model of typographic errors and their revisions. Specifically, our research questions are: 1) What process-based features might be indicative of typographic errors and their revision? 2) Is it possible to classify typographic error revisions in a copy task using process-based features? 3) Is it possible to transfer this model to a more natural writing task?
- **Methodology**: Three analyses were conducted. First, typographic errors and their revisions were characterized within copy tasks, where every error could be considered a typographic error. For this, process-based information, such as temporal and character bigram properties, were extracted from keystroke data from 2,103 copy tasks. Second, machine learning was used to create a process-based model on typographic error revisions to automatically identify these revisions within the copy task. Four different classification models were used: random forest, support vector machine, logistic regression, and naive Bayes. Lastly, this model was evaluated in a more natural setting: on keystroke data obtained from 66 regular (source-based) writing tasks. False positives and false negatives were analyzed to identify possible points of improvements for the model.
- **Results**: The characterization showed that typographic errors are made and revised in a variety of ways. However, we do see some general patterns in character bigram properties and timing of the keystrokes, which might be used to model typographic error revisions using process data only. Results on these process-based models indeed showed that it is possible to identify typographic errors using keystroke information only, especially in a copy task. Yet, the models on the regular (source-based) writing task still lead to a high number of false positives.
- **Discussion**: Even though no content information (such as word lists) was included, already a high performance could be reached on the models of typographic error revisions in the copy task. The lower performance on the source-based writing task could be explained by the fact that the data was unbalanced and the model was trained on the copy task.
- **Conclusion**: This study showed that process-based data, and specifically character bigram and temporal features, could be used to model typographic error revisions in copy tasks, and to a lesser extent in source-based writing tasks. Using these models, a more nuanced analysis of fluency and revision in writing can be performed.

Keywords: Classification, Revision, Typographic Error, Writing Analytics

1.0 Background

The analysis of writing, and especially writing processes, is complex. Writing is a non-linear activity, where processes such as planning, translating, and reviewing interact and can happen at any time during the writing process (Flower & Hayes, 1980). In most cases/writing studies, these cognitive processes have to be inferred indirectly, through observation. Therefore, it is hard to identify when these cognitive processes occur and how they interact. In addition, these processes are influenced by factors within the individual and the task environment (Hayes, 2012), such as the writing medium, language, working memory capacity, fluency, genre, audience, motivation, time, and knowledge on task (Lindgren & Sullivan, 2006b). To better understand writing, it is important to understand how these factors influence the writing process.

Currently, a large extent of writing is computer-based, using a word processor. It has been well-established that this medium has an influence on the writing process and writing product (Haas, 1989; Lindgren & Sullivan, 2019; Van Waes & Schellens, 2003). Even relatively simple processes or actions on this medium, such as the revision of a typing error (typo), can already have an influence on the writing process. These typing errors are numerous: An analysis of online writing by Grammarly showed we make on average 13.8 errors per 100 words in the morning and 17.0 errors per 100 words in the evening (Hertzberg, 2017).

For the analysis of writing, the importance of identifying typing errors and their revisions is twofold. On the one hand, they are low-level, and hence less-important types of revision, which would be beneficial to filter or analyze separately. Already, several studies have distinguished between different types of revision. One of the most common distinctions is surface revisions versus semantic or deep revisions (Faigley & Witte, 1981; Lindgren & Sullivan, 2006a). In addition, typing error revisions should be analyzed separately, as these are considered to reflect cognitively different actions (Wengelin, 2007). By treating typing errors separately, a more nuanced analysis of fluency and revision can be made (Barkaoui, 2016; Wengelin, 2007). For example, studies predicting writing quality often find contradicting results on the effect of the number of revisions on writing quality (e.g., Allen et al., 2016; Crawford, Lloyd, & Knoth, 2008; Roscoe, Jacovina, Allen, Johnson, & McNamara, 2016; Xu, 2018). This might be explained by the different types of revisions: One student might just be a careless typist who makes many typing errors, while another student is actually making a series of thoughtful revisions. On the other hand, typing errors, and especially the revision of typing errors, can (unwillingly) break the (linear) flow in writing. This can result in disfluency and activation of other subprocesses (Leijten, De Maeyer, & Van Waes, 2011; Lindgren & Sullivan, 2006b). To examine the influence of these typing errors on the writing process, it is necessary to be able to identify these errors first.

In addition to the importance for writing analytics, the identification and characterization of specific types of revisions during the writing process, such as typing error revisions, is of importance for writing instruction and feedback as well. First, the analyses of these revisions within multiple settings might shed light on effective writing strategies for dealing with these types of errors. For example, an effective writing strategy might be to not immediately revise every small typing error, as this might disrupt the flow in writing (cf. Leijten et al., 2011). In addition, the automatic identification of different types of revisions makes it possible and easier to observe, evaluate, and reflect on the effect of the used practice on revisions in multiple settings. This allows for evidence-based practices within writing instruction (Graham, 2019). Second, identification of revisions could provide important insights for the content of (automated) feedback on writing. For example, it has been argued that automated writing feedback should include information on students' revising behavior to better understand how revision affects their writing quality (Roscoe et al., 2016). Lastly, the writing process, and specifically revisions within the writing process, are evidently related to the writing product and hence, writing quality. The analysis and better understanding of these processes and their relation to writing quality (e.g., Allen et al., 2016; Zhu, Zhang, & Deane, 2019) play an important role in writing assessment. These insights into the writing process may be used for (automatic) assessment, based on process characteristics, either in combination with product characteristics or not.

Given the importance of the identification of revisions for both writing analytics and writing instruction, we aim to build a process-based model of typing errors and their revisions to automatically identify typing error revisions within the writing process, as opposed to other revisions. In this study, we specifically focus on typographic errors. Typographic errors are unintended keystrokes or slips of the finger leading to, for example, the transposition of two keys (e.g., *fro* instead of *for*). To model these errors, data obtained from keystroke logging software will be used. Keystroke logging has been used as a tool to gain insight into the writing process (Leijten & Van Waes, 2013; Lindgren & Sullivan, 2019). Several studies have used keystroke logs to manually classify revisions into typographic error revisions and other types of revisions (see e.g., New, 1999; Stevenson, Schoonen, & de Glopper, 2006). However, the automatic analysis of typographic error revisions using process data is still understudied.

2.0 Literature Review

The occurrence of typing errors has been well studied from a technical and ergonomical perspective in the fields of human computer interaction and information retrieval. Studies date back from the early 20th Century, when typewriters became commercially available (Kano, Read, Dix, & MacKenzie, 2007). An early review of typing errors already showed the variety of error categorizations that were made (Dvorak, Merrick, Dealey, & Ford, 1936). In addition, confusion matrices have been created for typewriter keys, showing for 60,000 errors the intended letter versus the actual typed letter (Lessenberry, 1928). Of these errors, 60% could be considered "adjacent" errors: The key was confused with an adjacent key on the keyboard (Kano et al.,

2007). Later, automatic classifications and automatic correction of typing errors became a key topic (see e.g., Peterson, 1980).

However, these studies cannot be directly used in the field of writing analytics to increase the understanding of the effect of typing errors on the writing process. There are two main issues: 1) typing errors are usually identified in the writing product only, not in the writing process; 2) no distinction is being made between different types or causes of typing errors. In this study we aim to address these two issues.

First, previous work generally identified typing errors within the writing product, but not the writing process. Therefore, only typing errors that were not corrected during the writing process and hence remained in the writing product were analyzed (Wobbrock & Myers, 2006). Especially with today's spell checkers, only a few typing errors will be left in the writing product. Thus, by analyzing typing errors within the writing product only, a large majority of typing errors are ignored. In addition, this analysis makes it impossible to determine the cause or the influence of typing errors, and especially typing error revisions, on the writing process and dynamics of writing. By analyzing typing errors during the writing process, we could gain evidence on both the (timing of the) production and correction of typing errors (Dhakal, Feit, Kristensson, & Oulasvirta, 2018; Wobbrock & Myers, 2006).

The identification of typing errors within the writing product is commonly done using a lexicon, comparing each misspelled word with the expected intended word. Here, the intended word usually is the closest word in the dictionary, based on some sort of edit distance, and possibly by taking into account the frequency of the word and the context of the word (see e.g., Damerau, 1964). However, the identification of typing errors using keystroke logging is not completely straight-forward. Within the writing process, typing errors are often made and corrected in half-written words, which can make it impossible to identify the intended word (Lindgren & Sullivan, 2019). Therefore, copy tasks have been used to identify typing errors during the writing process (see e.g., Dhakal et al., 2018; Wobbrock & Myers, 2006). In a copy task, participants have to transcribe a given text, and hence the intended word is known. However, for writing analytics we would like to identify typing errors in actual texts too, e.g., texts written by students. Therefore, we first build a process-based model of typing errors on data from a copy task, and thereafter test this model on a more natural task.

Second, previous studies typically do not distinguish between different types or causes of typing errors (Kano et al., 2007). After all, they commonly aim to identify and correct typing errors with the highest possible accuracy (Peterson, 1980), regardless of the type of error. Most of the early classifications of typing errors are based on edit operations. A large majority (80%) of the typos are caused by single letter errors: an extra letter (insertion), a missing letter (omission), a wrong letter (substitution), and a transposition of two adjacent letters (Damerau, 1964). The second most common type of typing errors are two-letter errors, including two extra letters, two missing letters, or two letters transposed around a third (e.g., *prodecure* instead of *procedure*) (Peterson, 1986). Kano and colleagues (2007) extended the classification with linguistic information, and identified omissions (letter, space, word, phrase), substitutions (letter,

next to letter, close to letter, capitalization, alternation, doubling, interchange, migration, word, phrase), transpositions (letter, word), insertions (letter, duplicated letter, next to error, close to error, space, duplicated space, symbol, function key, word, duplicated phrase), corrected no-errors (characters that were replaced with the same characters), and other errors (enter error, execution/hold key, unknown). By using keystroke data, Wobbrock and Myers (2006) identified five types of corrected and uncorrected errors: substitutions; insertions; omissions; no-errors; and non-recognition errors, key presses that did not result in actual characters produced (e.g., function keys).

However, in writing analytics, distinctions are made between different types of errors and revisions, as these are considered to reflect cognitively different actions which should be analyzed and interpreted separately (Wengelin, 2007). One specific type of typing error is the typographic error. Wengelin (2007, p. 73) defined typographic errors as "slips of the keyboard, i.e., errors that occur despite the writer's knowledge of how they are spelled." Slips can be seen as a human error where the action was not performed as it was intended (Norman, 1981). In Rumelhart and Norman's (1982) model of typing, based on the Activation-Trigger-Schema system, a typographic error occurs when a wrong keystroke schema is highly activated and the trigger conditions are met, resulting in the launching of the wrong keystroke. Even when the appropriate schema is activated, errors can be caused when schemas are triggered out of order or missed (Norman, 1981). For example, for transposition errors, the trigger conditions of the next keystroke are satisfied before the trigger conditions of the current keystroke, activating the next schema of the keystroke. Typographic errors are often opposed to orthographic errors or linguistic errors, that break the conventions of written language, such as errors in spelling, grammar, or punctuation (Lindgren & Sullivan, 2006a).

In this study, typographic error revisions will be identified using keystroke data. Keystroke data have shown to provide insight into the writing process (Leijten & Van Waes, 2013; Lindgren & Sullivan, 2019). Moreover, they have shown to distinguish different types of writing processes, such as text production and revision (Baaijen, Galbraith, & de Glopper, 2012). Keystroke data provide a fine-grained insight into how a text is constructed and reconstructed, allowing for a process-based and time-based analysis of revisions (Lindgren & Sullivan, 2006b). However, keystroke analysis has not yet been used to automatically identify typographic error revisions. Several studies did try to *manually* annotate revisions of typographic errors. It is considered especially difficult to distinguish typographic errors from orthographic errors; therefore, usually rules or guidelines were used for annotation. Wengelin (2007) indicated several properties of typographic errors:

> a) A typographic error can be a substitution of a letter, within a word, where the intended key is an adjacent key or a key with the same position for the other hand; or an omission of a letter.
> b) Typographic errors are rarely left in the final text and are usually corrected almost immediately.

c) Words with typographic errors are usually only corrected once and into the correct version.

Stevenson and colleagues provided a slightly more prescriptive description, indicating five possible cases which would be considered a typographic revision (Stevenson et al., 2006, pp. 230–231):

a) The pre-revision form does not conform to the orthographic rules of the language (e.g., *moore* 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 a present 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.
e) A letter is replaced by an adjacent letter on the keyboard.

In addition, if uncertainty remained, the timing of the revision was taken into account, where revisions made within one second from the previous keystroke were considered typographic revisions (Stevenson et al., 2006).

To summarize, these studies show that typing errors, and typographic errors specifically, can take many different forms, and have many different properties. However, there are also some common patterns that might be used for the automatic identification of typographic errors. Some of these patterns are related to semantic content and pronunciation, which might be hard to extract when the intended text is unknown. Other patterns might be easier to observe, such as the type of error, position of the characters typed on the keyboard, and the immediacy or timing (e.g., interkeystroke interval) of the revision. All these features might influence the probability of a typographic error.

3.0 Research Questions

In this study, we aim to build a process-based model of typographic errors and their revisions to automatically identify typographic error revisions within the writing process, as opposed to the writing product. Specifically, we aimed to answer three research questions:

1) What process-based features might be indicative of typographic errors and their revision?
2) Is it possible to classify typographic error revisions in a copy task using process-based features?
3) Is it possible to transfer this model of typographic error revisions to a more natural writing task?

4.0 Research Methodology

In order to address our research questions, three analyses were conducted using two different datasets. First, we characterized typographic errors and revisions in a dataset from a carefully manipulated copy task. In a copy task, all errors can be considered typographic errors since the correct spelling and grammar is provided to the writer. Likewise, all revisions can be considered typographic revisions. Accordingly, no manual annotation is needed, making it easier and less time-consuming to collect larger amounts of data about typographic errors and their revisions. Second, insight from this characterization was used to build a model of typographic errors on the copy task data. Two types of process-based features are included in the model: 1) temporal properties, focusing on the interkeystroke intervals of character bigrams preceding and following the typographic error; 2) character bigram properties, focusing on the frequency, adjacency, and keyboard position of the bigrams preceding and following the typographic error. Finally, this model was tested on data obtained from a more natural writing task (source-based synthesis) and evaluated using a manually annotated sample. In the following, we first discuss the two datasets collected and the cleaning and transformation of these data, followed by the three analyses conducted.

4.1 Copy Task Dataset

The copy task data has been collected from the Dutch copy task in Inputlog (Van Waes, Leijten, Pauwaert, & Van Horenbeeck, in press). This copy task has principally been designed to measure typing and motor skills in writing. The task is a strictly controlled task, which consists of seven parts with complementary levels of lexicality. In the task, participants were asked to repetitively copy: two characters alternatively, a sentence, four three-word combinations, and one set of blocks of consonants (non-words). Participants were instructed to transcribe as accurately and fast as possible. In total, data were available of 2,103 copy tasks conducted by 1,711 unique participants. The participants were all Dutch, with ages ranging from 8-83 years, with the majority between 15 and 25 years old (M = 23.6, SD = 12.6). Two-third of the participants were female (1,164, 68.0%).

For the current study, we only used the keystroke data from the word combination components in the copy task, as these were carefully constructed on the number of characters, word frequency, character bigram frequency, and mix of hand combinations. For an overview of the word combinations and their characteristics, see Table 1. The participants were asked to write each word combination seven times. In total, there were 59,423 attempts, consisting of 1,445,314 characters, of the word combinations. For every attempt and after every keystroke, the text transcribed so far (T) was computed. Every keystroke at the end of the attempt which did not belong to the attempt, but which was used as separation between two attempts, such as a space, comma, or period, was removed from the data, as this was considered not a part of the prompted text. Then, the edit distance between the transcribed text (T) and the prompted text (P) was calculated, using the restricted Damerau-Levenshtein distance (Boytsov, 2011). This metric calculates the minimum number of insertions, deletions, substitutions, or transpositions needed to

change the transcribed text into the prompted text (or vice versa). All attempts with less than two or more than 30 characters and all attempts where the final transcribed text had an edit distance larger than 90% of the number of characters in the prompted text, were considered non-serious attempts and hence removed. In total, 58,452 attempts remained for analysis.

Table 1

Bigram Properties of the Word Combinations in the Dutch Copy Task in Inputlog (Van Waes et al., in press)

		Number of bigrams							
Word combinations	Length	High freq	Low freq	Left-Left	Left-Right	Right-Left	Right-Right	Adja-cent	Repe-titive
vier mogelijke verbanden	23	0	20	4	5	4	4	7	0
drie belangrijke kinderen	23	0	19	4	4	3	4	7	0
vier duidelijke manieren	22	0	19	4	5	5	5	8	0
een chaotische cowboy	19	4	12	2	3	2	0	0	1

After data cleaning, typographic errors and revisions were coded in the keystroke log data file. Typographic errors were extracted largely following the procedure as described by Wobbrock and Myers (2006). First, the transcribed text was filled with dummies up to the length of the prompted text. A typographic error was flagged every time this distance between the filled transcribed text and prompted text did not decrease compared to the previous keystroke, i.e., when the transcribed text did not come closer to the prompted text. For the first character, a typographic error was flagged if the distance of the filled transcribed text was equal to the number of characters in the prompted text. A total of 46,996 typographic errors were identified (3.4% of the characters). Thus, an error was made in every 30 characters typed. A revision was defined as every single backspace or sequence of consecutive backspace keystrokes. If a revision removed a character labeled as typographic error, it was considered as a revision of that error. In 56% of the cases (26,411 errors) an error was revised. In total, there were 25,930 revisions, of which some revised multiple typographic errors. The rather low percentage of corrected errors could be explained by the specificity of the task, as participants were asked to type as fast and accurately as possible in the copy task.

4.2 Source-Based Writing Task Dataset

The source-based writing task (synthesis) dataset was a subset of the dataset collected in Leijten et al. (Leijten, Van Horenbeeck, & Van Waes, 2019; Leijten, Van Waes, Schrijver, Bernolet, & Vangehuchten, 2019). All the participants in this subset were also in the copy task dataset. The participants were asked to write a text of 200 to 250 words on humanitarian aid, renewable energy, climate change, or animal rights. Three sources on the given topic were provided: a

report, a web text, and a newspaper article. The participants got a maximum of 40 minutes to finish the assignment and were free to consult online tools or content on the internet. During the task, keystroke data were collected using Inputlog (Leijten & Van Waes, 2013). In total, data were available from 66 source-based writing tasks. The participants were all graduate students between 21-48 years (M = 27.4, SD = 8.1), and the majority (73%) was female. Similar to the copy task dataset, a revision was flagged for every sequence of backspaces. On average, the participants typed 2,980 characters (SD = 1,205) and made 116 revisions (SD = 62).

4.3 Analysis 1: Characterization of Typographic Error Revisions

For the characterization of typographic error revisions, several features were extracted from the copy task dataset, (partly) based on the current literature (see Literature Review section). In total four types of features were distinguished: type of typographic error, character bigram properties of the error, timing of the error, and revision of the error.

4.3.1 Error type. Each typographic error was classified as an insertion (addition), deletion (omission), substitution, or transposition. The type of error was inferred by comparing the transcribed text (T) and the expected text (X). X is a substring of the prompted text (P), with the same length as the transcribed text. In case of a deletion, a character was deleted from the expected text, and in case of an insertion, a dummy was added to the expected text.

4.3.2 Bigram properties. Bigram properties were extracted from character bigrams. In the remainder of this paper, whenever we mention bigrams, we refer to character bigrams. Bigram properties were extracted from the typed bigram in which the typographic error was made, the expected bigram in which the error was made (i.e., the bigram that should have been written), and the swapped bigram. For deletions, insertions, and substitutions, the swapped bigram was the combination of the typed and expected letter. For transpositions, the two swapped characters make up the swapped bigram. An example for every type of error and the corresponding bigrams can be found in Table 2. The type of error and bigram were extracted for every first typographic error within each attempt of transcribing the word combinations. Then, for every deletion and insertion that was not revised, the expected text (X) was realigned with the transcribed text, and the bigrams and error types were computed for the second typographic error within the attempt. This last step was repeated until this was calculated for every typographic error within each attempt.

Table 2

Examples of Possible Typographic Error Types with the Typed and Expected Text and Bigrams

	Word		Bigram		
Error type	typed	expected	typed	expected	swapped
Deletion	moglijke	mogelijke	gl	ge	le
Insertion	moogelijke	mogelijke	oo	og	og
Substitution	mofelijke	mogelijke	of	og	fg
Transposition	mogeiljke	mogelijke	ei	el	il

For each expected, typed, swapped bigram within a typographic error, four bigram characteristics were extracted: bigram frequency, repetitiveness, adjacency, and hand combinations. We did not include the actual bigrams, because the content will have a large effect on the bigrams typed and expected. The frequency of the bigrams was calculated as in (Van Waes, Leijten, Mariën, & Engelborghs, 2017), using the CELEX Lexical Database of the Dutch Centre for Lexical Information (Baayen, Piepenbrock, & van Rijn, 1993). The 30% most frequent bigrams in Dutch (e.g., le or ie) were classified as high-frequent, the 50% least frequent bigrams (e.g., ao or ow) as low-frequent. All others were defined as medium-frequent. The frequency of bigrams which included a space, start or end of sentence marker, were coded as missing. In addition, for each bigram we computed whether it consisted of a repetitive key, adjacent keys on the keyboard, or which hand combination would be used in case of touch typists (left-right, left-left, right-left, right-right, or unknown for bigrams which included a space or a middle key).

4.3.3 Timing. Interkeystroke intervals (IKI), the time from key onset to key onset, were extracted for the typographic error itself, the four keystrokes preceding the error, and the five keystrokes following the typographic error. Because the IKI length is highly influenced by the participant, we subtracted the mean IKI per participant from every IKI, resulting in the difference in IKI from the mean IKI.

4.3.4 Revision. For every revision, we indicated which typographic error(s) it revised. In addition, we extracted the number of backspaces, as well as the number of characters removed. Note, the number of backspaces can be unequal to the number of characters removed, for example when the number of backspaces is larger than the current number of characters transcribed. Lastly, we extracted the revision delay, the number of characters typed until the revision started, and the revision overflow: the number of characters additionally deleted after revising the typographic error. For example, if you would type form<<<<from, the revision delay is 2, because the error started with the mistyped o. The revision overflow is 1, since one additional character is (unnecessarily) revised (f).

For these four different types of features, descriptive statistics are reported and visualized. The four types of features are not independent. For example, the revision delay influences the

IKI timings following the typographic error. Therefore, the characteristics are not only discussed for each feature type individually, but also in relation to each other.

4.4 Analysis 2: Automatic Classification of Typographic Revisions in a Copy Task

To build a process-based model on typographic error revisions, we first built a model using keystroke logging data derived from the copy task. As all typographic errors and revisions are flagged within a copy task, no manual annotation is needed.

However, a process-based model on typographic revisions built on a copy task will basically learn whether a typographic error revision is made within the copy task, as all revisions in the copy task are considered typographic error revisions. This model will generalize badly to other, more natural tasks, as in those tasks revisions other than typographic errors will be present (e.g., spelling or wording revisions). In this case, the model will have learned to flag all revisions as typographic revisions. Therefore, we constructed a model on typographic *errors*, rather than typographic revisions. By using information on the keystrokes following the typographic error, we could subsequently identify whether the error was revised. Hence, we classified typographic error revisions using a two-step approach, by 1) classifying the typographic error, and 2) determining whether the typographic error was revised. First, this model is built on the copy task. Thereafter, the model trained on the copy task is evaluated on a more natural writing task: a source-based writing task.

For the classification of typographic errors, the cleaned and enriched dataset from the previous analysis was used, but only copy task data were included from participants who also completed the source-based writing task. This subset of the copy task dataset consisted of 73,795 characters typed of which 2,225 (3.0%) were typographic errors. The feature extraction was also based on the feature extraction from the previous analysis. However, not all these features could be extracted from the source-based writing task. For example, for the type of error or the swapped bigram, the intended word is needed, which is not present in natural writing tasks. Therefore, we only included features which could be extracted from both the copy task and the source-based writing task dataset. In total, three types of features were included.

4.4.1 Bigram properties. Bigram properties were collected from the last two characters at the cursor location. Although very uncommon in the copy task, in the source-based writing task participants often moved between different parts of the text, using a mouse or arrow keys. Hence, the bigrams were not extracted from the last two keystrokes typed, but from the last two characters at the cursor location. For every bigram, the current bigram and the four preceding bigrams were extracted (features up to error), as well as the five following bigrams. The former could be seen as the typed bigram and the latter as the expected bigram in Table 2. As in the characterization, bigram frequency, hand combination, key adjacency, and repetitiveness were extracted.

4.4.2 Timings. The timing of the keystrokes surrounding the typographic error were calculated using the IKI of these keystrokes. Again, the timings were grouped into the IKI of the

current keystroke and the four preceding keystrokes, as well as the IKI of the five following keystrokes.

4.4.3 Participant. Participant was included, as typographic errors are made differently across typists.

A model was trained on these features to classify for every keystroke whether or not it was a typographic error. The data was highly imbalanced; only 2.7% of the keystrokes constituted a typographic error. Therefore, random down sampling, from the caret package in R (Kuhn, 2018) was used to balance the data. With down sampling, the data set is randomly sampled so that all classes (normal keystrokes and typographic errors) have the same frequency as the minority class (typographic errors). In total, 2,038 typographic errors and 2,038 non-typographic errors were used to train the model. One hot encoding was used to transform the categorical bigram properties into binary features. Centering and scaling were applied to the timing features.

Two sets of the features were used: 1) features which only contained information up to the bigram or typographic error, and 2) features that contained both information from before and after the bigram or typographic error. Four classification models were trained on both feature sets: random forest, support vector machines with radial kernel, logistic regression, and naive Bayes. Random forests were chosen because they generally work well on categorical data. In addition, support vector machines were chosen, because they generally work well with sparse data. Lastly, logistic regression and naive Bayes were added to determine whether a simple model would work equally well. All models were implemented using the caret package in R (Kuhn, 2018), were trained on the F-score, and run using 10-fold cross-validation. The results of the classification of typographic errors were evaluated on the copy task data. As evaluation metrics, precision, recall, and F-score are reported. These metrics are identified below and shown in Figure 1:

$$\text{precision} = \frac{true\ positives}{true\ positives + false\ positives}, \text{ recall} = \frac{true\ positives}{true\ positives + false\ negatives},$$

$$\text{F-score} = \left(2 \cdot \frac{precision \cdot recall}{precision + recall}\right)$$

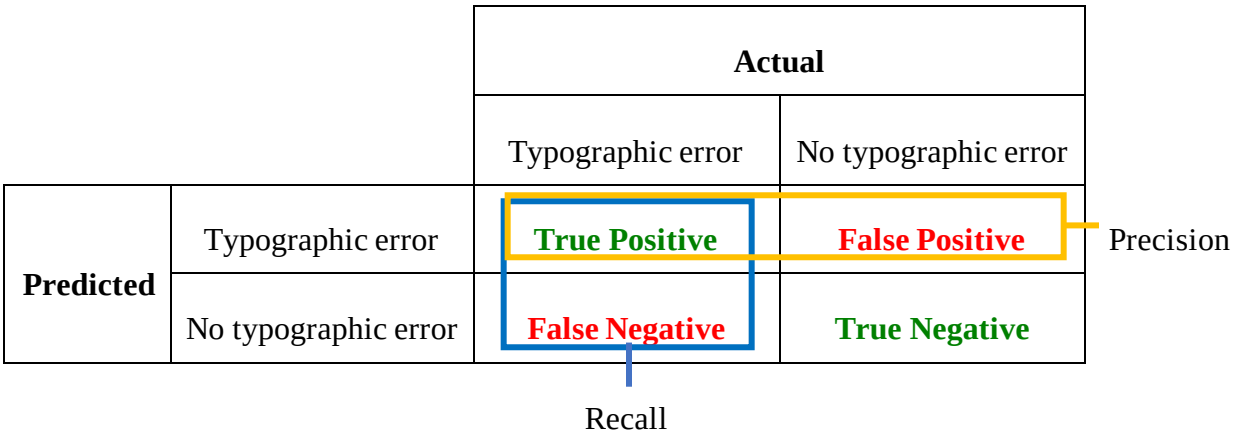

Figure 1. Confusion matrix and evaluation metrics for the classification of typographic errors.

4.5 Analysis 3: Automatic Classification of Typographic Revisions in a Source-Based Writing Task

For the classification of typographic error revisions on the source-based writing task, the same features were extracted as in Analysis 2. For the evaluation of the model on the source-based writing task dataset, a small subsample of six participants was manually annotated. For every revision, two human annotators identified whether it was a typographic error revision or not. If it was unclear, the revision was annotated with a question mark (see Lindgren & Sullivan, 2019). In total, 879 (15%) revisions were annotated. An inter-rater reliability of 88% was reached. Disagreements were resolved through discussion. If no agreement could be reached, or if it was still unclear whether it was a typographic error revision, the revision was annotated with a question mark. In total, 67 (7.6%) of the revisions were annotated with a question mark. For example, one participant typed *care f<about* (translated). Here, it is hard to identify whether this was a typographic error revision, because it is unclear whether the participant mistyped the *a*, or first wanted to start the word with an *f*, e.g., *for*, and later decided to use another word (*about*).

The obtained models from the classification of typographic errors in the copy tasks (Analysis 2) were in turn tested on the source-based writing task dataset. The model was trained on down-sampled data of the copy task, where there are equal amounts of typographic errors and non-typographic errors. In contrast, the source-based writing task dataset cannot be down sampled, as this would require data labeled with typographic errors, while this is actually the class we are trying to predict. Thus, the proportion of typographic errors in all keystrokes in the source-based writing task will be low. Therefore, testing this model immediately on all keystroke data of the source-based writing task would result in many false positives. The characterization of typographic errors showed that typographic errors are in general revised within five characters from the error. In addition, a revision itself can never be a typographic error. Therefore, the

model was only tested on the five keystrokes preceding every revision, resulting in a dataset of 25,502 keystrokes.

After testing the model on the source-based writing task, we still needed to determine whether the typographic error was revised. We flagged a revision as a typographic revision if the last backspace revised a typographic error. For example, if the revision contained three backspaces (e.g., *hoise<<<use* to type *house*), this would be considered as a revision with a delay of two keystrokes (*se*), if the keystroke revised by the last backspace, i.e., the third keystroke preceding the revision (the *i* in *ho*i*se*), was flagged as a typographic error. Note, we assume here that the revision overflow is always equal to zero (which was true for 95% of the cases in the copy task). In addition, all revisions above five keystrokes were identified as non-typographic error revisions. The flagged typographic error revisions were evaluated using the manual annotations.

In addition, false positives and false negatives were analyzed to gain insight into possible points of improvements for the model. To analyze the errors, the keystroke log surrounding the wrongly identified typographic revision or missed typographic revision was manually inspected. The false positives and false negatives were then grouped into error categories to ease the interpretation of the errors. In addition, this could be used to identify where the largest gains in the model could be reached. For example, the categories with the largest number of false positives or false negatives, and which could be easily solved, should be prioritized in the next iteration of the model.

5.0 Results

5.1 Characterization of Typographic Error Revisions

5.1.1 Error type. In total, 46,996 typographic errors were identified, of which most (57%) were substitutions, followed by insertions (18%), transpositions (13%), and deletions (11%). Table 3 presents an overview of how often each type of error was revised. Substitutions were revised most, followed by transpositions. Deletions were least often revised.

Table 3

Number of Revised and Non-Revised Typographic Error Types

	Deletion	Insertion	Substitution	Transposition
Non-revised	3,407	5,284	9,047	2,847
Revised	1,871	3,261	17,950	3,329
Total	5,278	8,545	26,997	6,176

5.1.2 Bigram properties. Most typographic errors were made when low-frequent bigrams were *expected*: In 7.2% of the cases where a low-frequent bigram had to be transcribed, a typographic error was made, as opposed to 3.1% of the high-frequent bigrams (see Figure 2).

Typographic errors were least common when repetitive bigrams were expected (1.3%). However, there was only one repetitive bigram in the expected text (*ee*); thus, this repetition could also be caused by the high frequency of this bigram in Dutch. In contrast to the *expected* bigrams, repetitive bigrams were most often (wrongly) *typed* (8.6%), followed by low-frequent bigrams (8.1%), as opposed to high-frequent bigrams (2.3%). Adjacent keys were least often wrongly typed (2.1%). No clear relation was found between the hand combination and the proportion of typographic errors in the typed and expected bigrams.

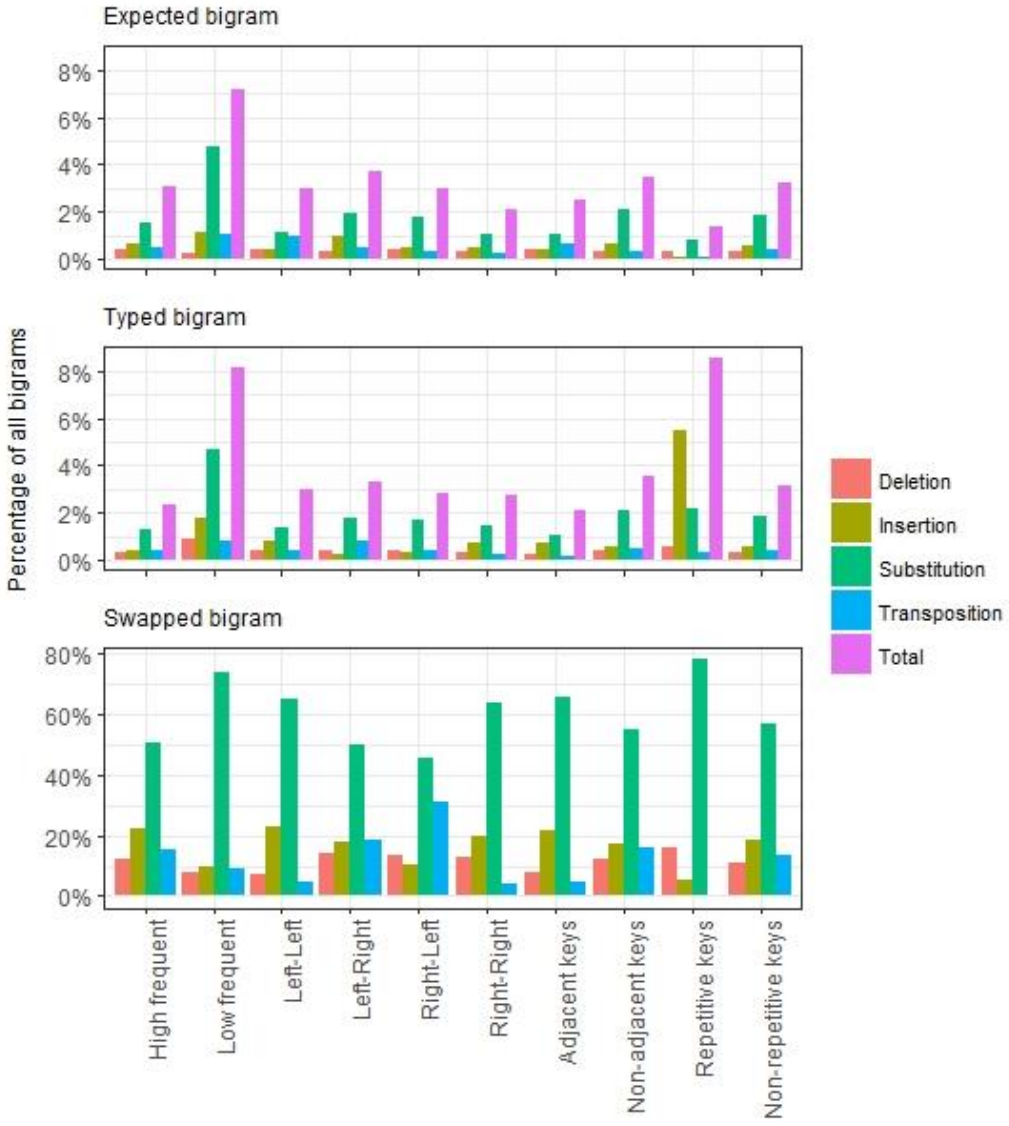

Figure 2. The percentages of errors within each bigram feature for expected and typed bigrams.

The proportions of typographic errors for the specific bigram properties were also analyzed for each type of error. Most types of errors did not show a clear pattern in which expected or typed bigram properties frequently occurred. Only for substitutions and insertions a pattern emerged: Substitutions were most common when a low-frequent bigram was *expected*. Insertions were most common when repetitive keys were *typed* (5.5%). These insertions could be the case where someone accidentally presses a key one additional time, e.g., *moore* instead of *more*. Insertions were fairly uncommon in alternating hand combinations (Left-Right, Right-Left), compared to same hand combinations (Right-Right, Left-Left).

The bigram properties for the different types of error in the swapped bigrams showed a somewhat more distinctive pattern (see Figure 2). Since the swapped bigrams are only available for bigrams in which typographic errors occur, all four error types add up to 100%. For bigram frequency in the swapped bigrams, insertions were more common when the swapped bigram was high-frequent (22%), compared to low-frequent (10%), while substitutions were more common in low-frequent bigrams (74%), compared to high-frequent (51%). For the hand combinations, substitutions were more common in same hand combinations, compared to alternating hand combination, while transpositions were more common in alternating hand combinations compared to same hand combinations. Adjacent keys were often substituted, but not transposed. Repetitive keys were commonly substitutions or deletions, where deletions indicate that two repetitive keys where prompted, but only one was typed (e.g., *ber* instead of *beer*).

5.1.3 Timing. The timings of the keystrokes around the typographic error showed a relatively clear pattern for the error (see Figure 3). The IKI of a revised typographic error was on average 46ms longer than the mean IKI of the participant. Thus, typists slow down when making a typographic error. Yet, the variance was large (*SD* = 170ms), indicating that this effect varies across errors. The IKIs of the keystrokes preceding typographic errors tended to increase. Interestingly, even when the typographic errors were not revised, the IKI still increased, up to an IKI of 37ms longer than the mean IKI, at the keystroke of the typographic error. Again, the variance was large (*SD* = 203ms).

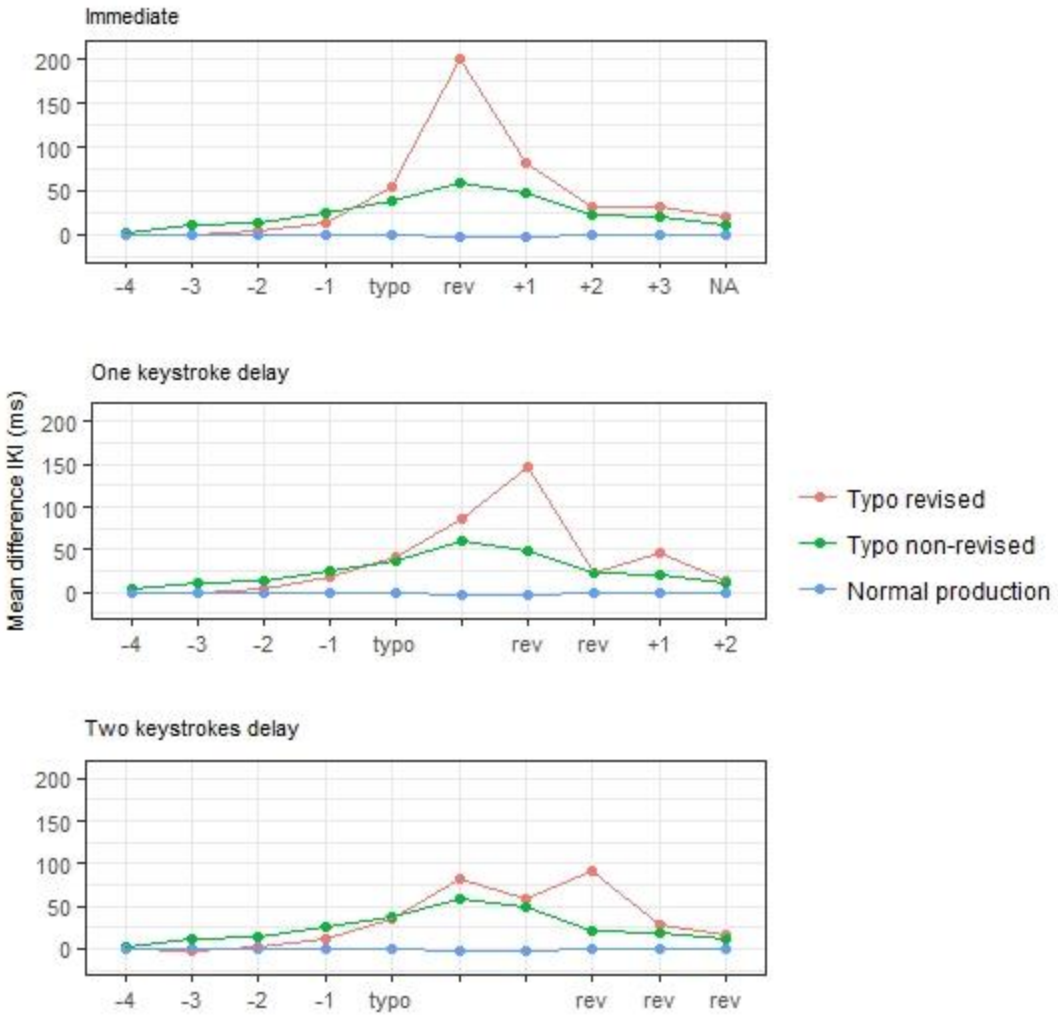

Figure 3. The interkeystroke interval (IKI) before and after (revised) typographic errors, compared to normal production.

After the error, the pattern of IKIs highly depended on how and when the typographic error was revised. When the error was not revised, the IKI increased for one keystroke directly after the error (M = 59ms, SD = 276ms above mean IKI) and then slowly decreased towards the mean IKI. When the error was immediately revised, a large increase could be found in the IKI immediately after the error (M = 201ms, SD = 251ms), probably indicating the time needed to move the hand towards the backspace key. When the error was revised later, e.g., delayed by a few keystrokes, there was still an increase directly after the error, but with the peak IKI belonging to the actual revision.

5.1.4 Revision. In total, 25,930 revisions were identified, of which 23,817 (92%) revised a typographic error. In the other cases, character(s) were replaced with the same character(s), so-called no-errors (Wobbrock & Myers, 2006). A large majority of revisions only revised a single

typographic error (84%), but sometimes two (6%), three (1%), or more than three errors (0.6%) were revised. Only 56% of the revisions revised a typographic error immediately after it was made. In all other cases, the revision was delayed with one keystroke (25%), two keystrokes (9%), three keystrokes (5%), or more than three keystrokes (5%). Sometimes the revision was even delayed with more than 10 keystrokes (0.1%). The latter indicated that the revision was made over word boundaries. Lastly, in the current dataset, it was relatively uncommon to revise more keys than necessary: There was a revision overflow of one letter in only 4%, and in 0.6% the overflow consisted of more than one letter.

To summarize, the characterization showed that typographic errors are made and revised in a variety of ways. However, we do see some patterns in the process data which might be used to model typographic error revisions using process data only. For example, the bigram properties indicated that typographic errors are more common in places where a low-frequent bigram is expected, and substitutions are common between adjacent keys for which the same hand is needed. Additionally, the IKI is increased compared to the mean IKI when a typographic error is made and increases even further when the IKI is revised. Therefore, in the analysis study we classify typographic errors within writing tasks, using process variables described above.

5.2 Automatic Classification of Typographic Revisions in a Copy Task

The models for the prediction of typographic errors were first evaluated on the copy task dataset. The models including only features up to the typographic error all exceeded the baseline accuracy (majority class) of 0.5 (see Table 4). Random forests were overall the best model, with a precision of 0.747, recall of 0.706, and an F-score of 0.726. In contrast, the models using both information from before and after the typographic error performed much better. Again, random forests were the best model, with a precision of 0.857, recall of 0.863, and an F-score of 0.860. The support vector machine was only slightly better in recall, with a recall of 0.872. Thus, process-based data can be used to predict a typographic error with a relatively high performance.

Table 4

Performance of the Prediction of Typographic Errors in the Copy Task Dataset

	All features, M(SD)			Only features up to error, M(SD)		
Model	Precision	Recall	F-score	Precision	Recall	F-score
Random forest	**.857(.02)**	.863(.01)	**.860(.01)**	**.747(.02)**	**.706(.04)**	**.726(.02)**
Support vector machine	.815(.02)	**.872(.03)**	.843(.02)	.669(.03)	.668(.04)	.669(.03)
Logistic regression	.746(.01)	.790(.03)	.767(.02)	.655(.02)	.658(.04)	.656(.02)
Naive Bayes	.637(.03)	.778(.03)	.700(.03)	.623(.01)	.524(.03)	.569(.02)

Note. Majority class baseline accuracy is .5.

5.3 Automatic Classification of Typographic Revisions in a Source-Based Writing Task

After the evaluation of the models of typographic revisions in the copy task, the models were tested on the source-based writing task. From the prediction of typographic errors, the corresponding revision was identified as revising a typographic error or not. The number of revisions in the source-based writing task classified as typographic error revision varied somewhat for the different models (see Table 5). For the models using only features up to the typographic error, around 50% of the revisions were classified as a typographic error revision. The models using all features were greedier, with up to 71% of the revisions classified as typographic error revisions for the random forest model.

Table 5

Number of Revisions Predicted as Typographic Revision in the Source-Based Writing Task Dataset

Model	All features	Only features up to error
Random forest	4,154 (71%)	3,122 (54%)
Support vector machine	3,642 (62%)	2,364 (41%)
Logistic regression	2,908 (50%)	2,560 (44%)
Naive Bayes	3,178 (54%)	2,775 (48%)

The models were evaluated with the annotated sample of the source-based writing task. In total, 341 revisions were annotated as a typographic error revision (42%). Hence, a majority class predictor would result in an accuracy of 58%, which could be seen as the lower baseline. The inter-rater agreement of 88% could be seen as an upper baseline. All models which only included the features up to the typographic error did not outperform the lower baseline (see Table 6). The models including all features showed better results. Again, random forests proved to be the best model. The accuracy, with 59%, only outperformed the lower baseline with 1%. This low accuracy was mostly caused by the low precision, indicating a large number of false positives. However, the recall was quite high: 79% of the all the typographic errors were retrieved. Interestingly, the revisions which could not be annotated by humans (indicated with a question mark) were mostly modeled as typographic error revisions by the machine: from 72% coded as typographic error revision by the naive Bayes model up to 97% by the random forest model.

Table 6

Accuracy of the Prediction of Typographic Revision in the Source-Based Writing Task Dataset

	All features			Only features up to error		
Model	Precision	Recall	F-score	Precision	Recall	F-score
Random forest	**.51**	**.79**	**.62**	**.49**	**.56**	**.52**
Support vector machine	.50	.67	.57	**.49**	.49	.49
Logistic regression	.45	.48	.47	.47	.48	.48
Naive Bayes	.48	.52	.50	.48	.55	.51

Note. Lower baseline accuracy (majority class) is .58. Upper baseline accuracy (inter-rater agreement) is .88.

The false positives and false negatives of the models with all features were further explored to provide insight in possible model improvements. In 140 (17%) cases, all models wrongly predicted a typographic error revision (false positive). Three common themes were found in the false positives. First, the false positives consisted of many "no-errors," where character(s) were replaced by the same character(s). Second, some of the false positives included revisions of punctuation markers, for example, changing a space into a comma, followed by a space. Lastly, sometimes a whole word was replaced by another word. In 47 (6%) cases, all models wrongly predicted a non-typographic error revision (false negative). Most false negatives occurred in typographic error revisions that occurred quickly after a failed attempt to revise a typographic error. In addition, false negatives occurred in transpositions, when the error was at the start of the word or when the error was only revised after the word was finished (including the space after the word).

6.0 Discussion

In this study we aimed to build a process-based model of typographic errors (slips of the fingers) and their revisions to automatically identify typographic error revisions within the writing process, as opposed to the writing product. For this, three different analyses were run. First, typographic errors were characterized on the type of error, bigram properties, timing, and revision, using data from a copy task. In line with previous studies that characterized typing errors in general (Dhakal et al., 2018; Wobbrock & Myers, 2006), substitutions were found as the most common typographic errors. Substitutions were also most often revised. Typographic errors were mostly revised within a few characters. This verifies one of the guidelines for manual annotation of typographic errors: Typographic errors are usually corrected almost immediately (Wengelin, 2007).

Typographic errors were most common when a low-frequent bigram was *expected*, as opposed to high-frequent bigram. Likewise, the wrongly *typed* bigrams were also more often low-frequent bigrams than high-frequent bigrams. *Swapped* bigrams, often presented in confusion matrices of typing errors (see e.g., Kernighan, Church, & Gale, 1990), provided

evidence for the relation between the position of the key on the keyboard and the error. For example, substitutions were more common in same hand combinations, while transpositions were more common in alternating hand combinations. These findings can be explained by Rumelhart and Norman's (1982) model of typing. According to their model, transposition errors can only occur if the wrong schema is triggered before the correct schema. This error would be more likely in alternating hands, because the fingers on the other hand have a speed advantage, as they are less constrained by the motions of the other fingers of the first hand (Rumelhart & Norman, 1982). In addition, transpositions would be more likely in adjacent keys from the same hand, as the palm helps rather than hinders movement towards the next finger. However, the later was not confirmed in our characterization: Adjacent keys were more often substituted than non-adjacent keys, but not transposed. A possible explanation for this might be that the writers in the copy task did not all type with ten fingers. When typing with two fingers, there is no speed advantage for adjacent keys (while there still is one for alternating hands). Hence, transpositions in adjacent keys might be less plausible when the writer is not a touch typist. However, the typing skill of the participants was not collected in the current study, thus this hypothesis could not be tested.

The timing of the keystrokes before the typographic error showed a slight increase in IKI (above the mean IKI). This might indicate that the error was caught prior to when it was made, i.e., the error was anticipated, but with insufficient time to prevent the error (Norman, 1981). The timings after the error also showed an increase in IKI, with a peak at the first backspace. Interestingly, even when the error was not revised, an increase could be found in the IKI after the error. This result might be related to the fact that the writer did notice the error; the correct schema is activated, but is not triggered yet. Eventually, the schema loses activation (e.g., due to decay) and no revision is made (cf. Norman, 1981).

Second, based on insights gained from the characterization of typographic errors, a process-based model was built on typographic errors in a copy task. Even though no content information (such as word lists) was included, already a high performance could be reached. Participant, timing of keystrokes, and bigram properties before and after a typographic error could already predict a typographic error with an F-score of 0.86 (random forest). The models were in turn tested on a natural writing task: a source-based writing task. Again, random forests was the best model, but it only slightly outperformed the majority class baseline. The model was rather greedy, with a high recall, yet had relatively low precision.

The low performance on the source-based writing task, and especially the large number of false positives, might be explained by limitations in the current study. First, the models were trained on a copy task, which is substantially different from a source-based writing task. Although the correction of typographic errors might be considered as a relatively consistent motor process within participants, we found that the characteristics might differ across tasks. For example, typographic errors in the copy task were revised fairly quickly, but errors in the source-based writing task were sometimes only revised after more than ten characters. Thus, the model might be further improved by training on the source-based writing task. However, this was not

the purpose of the current study; we tried to model the errors with data from a copy task, such that time-intensive manual annotation would not be necessary. Second, the model was trained on down-sampled, and consequently, balanced data, while the model was tested on unbalanced data, even though we did make the test set slightly more balanced by extracting only the five keystrokes preceding and following a revision. The testing on the unbalanced dataset might explain the high recall or the greediness of the model and hence the large number of false positives. Third, currently only process-related features were included in the model. Additional features, such as content information might be used to improve the model.

7.0 Conclusion

To conclude, this study provides significant insight into the dynamics of typographic error production and revision in online writing processes. It was shown that especially temporal and bigram features are indicative of typographic errors and their revisions. These properties were found to be useful to model typographic error revisions in a copy task, but the model transferred less well to a more natural task: a source-based writing task. Although the model is not yet perfect, a first step is made into the direction of the automatic typographic error revisions classification, using keystroke data. This classification may be used for more nuanced analyses of fluency and revision, as in this way typographic errors could be analyzed separately. In addition, this classification could be used to determine their influence on the writing process, e.g., triggering reviewing episodes. Lastly, this classification might be used in writing instruction and feedback, for example, by automatically assessing the revision process.

8.0 Directions for Further Research

Possible directions for further research include the improvement of the classification of typographic error revisions in natural writing tasks. First, future work should identify whether another way of balancing the training data—keeping it more truthful to the actual proportions in the dataset—or smarter ways of balancing the test data would result in better performance. One solution for the latter would be to only extract keystrokes from a revision event, where a *revision event* is defined as all keystrokes that are removed by the revision and all keystrokes that are replaced.

Second, future work should identify how the model could be further improved. Additional process-based features could be included, such as information on character position (e.g., word initial bigrams, cf. Crump & Logan, 2010) or dynamics in writing fluency (cf. Van Waes & Leijten, 2015) or bursts (cf. Baaijen & Galbraith, 2018). In addition, content information such as word lists or information about the writer, such as typing skill (touch typists versus hunt-and-peck typists) or language proficiency, might increase the accuracy of the model.

The exploration of the false positives and false negatives showed additional possible points of improvement for the model. Clustering of the errors resulted in some groups that might be easily addressed in the model and hence could be prioritized. For example, revisions including punctuation marks often led to false positives. This might be because key adjacency was not

coded for punctuation keys. Thus, the model could be improved by including key adjacency for punctuation. In addition, the copy task did not include punctuation, hence training the model on a copy task including full sentences with punctuation might translate better to the source-based writing task. In addition, revisions at the start of a word or after a word had finished often led to false negatives. Therefore, it might be useful to also include information on the bigram properties at the start or end of a word or sentence.

No-errors were the most common false positives, indicating that the largest improvement gain might be reached when this error is addressed. No-errors are revisions where the revised characters are replaced by the same characters (Wobbrock & Myers, 2006). No-errors were also common in the copy task, indicating that they actually might be some sort of typographic error (e.g., when you mistakenly thought you made an error). However, no-errors could also indicate a different cognitive process, where you consider writing a different word or word-form, but then decide to stick with the first word. While the former intuitively might be considered a typographic error, the latter is not. With the current human annotation (based on the keystroke log), this nuance cannot be distinguished. Future work should further investigate these errors using thinking-aloud, stimulated recall, or eye tracking with touch typists, to uncover the intention of the writer and identify whether this could be automatically classified (Lindgren & Sullivan, 2006a, 2019). Lastly, it would be interesting to implement this model in keystroke logging software to make it possible for researchers to identify typographic error revisions and to analyze these separately.

Note

This research was presented at the Eighth International Conference on Writing Analytics, Winterthur, Switzerland, September 2019.

Author Biographies

Rianne Conijn is a PhD candidate in the department of Cognitive Science and Artificial Intelligence at Tilburg University, the Netherlands. Her research interests include the analysis and interpretation of (sequences of) online and offline learning behavior to improve learning and teaching.

Menno van Zaanen is a Professor in Digital Humanities at the South African Centre for Digital Humanities (SADiLaR), South Africa. He has a background in computational linguistics. His areas of research include multi-modal structuring of data, multi-modal information retrieval, and applying digital techniques in the humanities.

Mariëlle Leijten is an Associate Professor in Professional Communication at the University of Antwerp, Belgium. She has been involved in different types of writing research, with a current focus on source-based writing processes.

Luuk Van Waes is a Professor in Professional Communication at the University of Antwerp, Belgium. He has been involved in different types of writing research, with a focus on digital media and (professional) writing processes.

References

Allen, L. K., Jacovina, M. E., Dascalu, M., Roscoe, R. D., Kent, K., Likens, A. D., & McNamara, D. S. (2016). {ENTER}ing the time series {SPACE}: Uncovering the writing process through keystroke analyses. *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*, 22–29. Retrieved from https://eric.ed.gov/?id=ED592674

Baaijen, V. M., & Galbraith, D. (2018). Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction*, *36*(3), 199–223. https://doi.org/10.1080/07370008.2018.1456431

Baaijen, V. M., Galbraith, D., & de Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, *29*(3), 246–277. https://doi.org/10.1177/0741088312451108

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical data base on CD-ROM*.

Barkaoui, K. (2016). What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal*, *100*(1), 320–340. Retrieved from https://doi.org/10.1111/modl.12316

Boytsov, L. (2011). Indexing methods for approximate dictionary searching: Comparative analysis. *Journal of Experimental Algorithmics (JEA)*, *16*, 1–1. Retrieved from https://doi.org/10.1145/1963190.1963191

Crawford, L., Lloyd, S., & Knoth, K. (2008). Analysis of student revisions on a state writing test. *Assessment for Effective Intervention*, *33*(2), 108–119. Retrieved from https://doi.org/10.1177/1534508407311403

Crump, M. J., & Logan, G. D. (2010). Hierarchical control and skilled typing: Evidence for word-level control over the execution of individual keystrokes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1369.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176. Retrieved from https://doi.org/10.1145/363958.363994

Dhakal, V., Feit, A. M., Kristensson, P. O., & Oulasvirta, A. (2018). Observations on typing from 136 million keystrokes. *Conference on Human Factors in Computing Systems-Proceedings*, *2018*. Retrieved from https://doi.org/10.1145/3173574.3174220

Dvorak, A., Merrick, N. L., Dealey, W. L., & Ford, G. C. (1936). *Typewriting behavior*. New York: American Book Company.

Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication*, *32*(4), 400–414.

Flower, L., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication*, *31*(1), 21–32. Retrieved from https://doi.org/10.2307/356630

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, *43*(1), 277–303. https://doi.org/10.3102/0091732X18821125

Haas, C. (1989). Does the medium make a difference? Two studies of writing with pen and paper and with computers. *Human-Computer Interaction*, *4*(2), 149–169.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*(3), 369–388.

Hertzberg, K. (2017, September). Grammarly analysis shows we write better by day [Blog post]. Retrieved from https://www.grammarly.com/blog/analysis-shows-we-write-better-day/

Kano, A., Read, J. C., Dix, A., & MacKenzie, I. S. (2007). ExpECT: An expanded error categorisation method for text input. *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but Not as We Know It-Volume 1*, 147–156. Retrieved from https://dl.acm.org/citation.cfm?id=1531314

Kernighan, M. D., Church, K. W., & Gale, W. A. (1990). A spelling correction program based on a noisy channel model. *Proceedings of the 13th Conference on Computational Linguistics-Volume 2*, 205–210. Retrieved from https://doi.org/10.3115/997939.997975

Kuhn, M. (2018). *caret: Classification and Regression Training*. Retrieved from https://CRAN.R-project.org/package=caret

Leijten, M., De Maeyer, S., & Van Waes, L. (2011). Coordinating sentence composition with error correction: A multilevel analysis. *Journal of Writing Research*, *2*(3), 331–363. https://doi.org/10.17239/jowr-2011.02.03.3

Leijten, M., Van Horenbeeck, E., & Van Waes, L. (2019). Analysing keystroke logging data from a linguistic perspective. In E. Lindgren & K. Sullivan (Eds.), *Observing writing: Insights from keystroke logging and handwriting* (pp. 71–95). Retrieved from https://doi.org/10.1163/9789004392526_005

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Leijten, M., Van Waes, L., Schrijver, I., Bernolet, S., & Vangehuchten, L. (2019). Mapping master's students' use of external sources in source-based writing in L1 and L2. *Studies in Second Language Acquisition, 41*(3), 555–582. https://doi.org/10.1017/S0272263119000251

Lessenberry, D. D. (1928). *Analysis of errors*. Syracuse, NY: LC Smith and Corona Typewriters.

Lindgren, E., & Sullivan, K. P. (2006a). Analysing online revision. In *Computer keystroke logging and writing: Methods and applications* (pp. 157–188). Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A143735

Lindgren, E., & Sullivan, K. P. (2006b). Writing and the analysis of revision: An overview. In *Computer keystroke logging and writing: Methods and applications* (pp. 31–44). Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A152837

Lindgren, E., & Sullivan, K. P. (2019). *Observing writing: Insights from keystroke logging and handwriting*. Retrieved from https://doi.org/10.1163/9789004392526

New, E. (1999). Computer-aided writing in French as a foreign language: A qualitative and quantitative look at the process of revision. *The Modern Language Journal*, *83*(1), 81–97. Retrieved from https://doi.org/10.1111/0026-7902.00007

Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88*(1), 1. https://doi.org/10.1037//0033-295X.88.1.1

Peterson, J. L. (1980). Computer programs for detecting and correcting spelling errors. *Communications of the ACM, 23*(12), 676–687. Retrieved from https://doi.org/10.1145/359038.359041

Peterson, J. L. (1986). A note on undetected typing errors. *Communications of the ACM, 29*(7), 633–637. Retrieved from https://doi.org/10.1145/6138.6146

Roscoe, R. D., Jacovina, M. E., Allen, L. K., Johnson, A. C., & McNamara, D. S. (2016). Toward revision-sensitive feedback in automated writing evaluation. *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*, 628–629. Retrieved from https://eric.ed.gov/?id=ED586437

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science*, *6*(1), 1–36. Retrieved from https://doi.org/10.1016/S0364-0213(82)80004-9

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*, *15*(3), 201–233. Retrieved from https://doi.org/10.1016/j.jslw.2006.06.002

Van Waes, L., & Leijten, M. (2015). Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers and Composition*, *38*, 79–95. https://doi.org/10.1016/j.compcom.2015.09.012

Van Waes, L., Leijten, M., Mariën, P., & Engelborghs, S. (2017). Typing competencies in Alzheimer's disease: An exploration of copy tasks. *Computers in Human Behavior*, *73*, 311–319. https://doi.org/10.1016/j.chb.2017.03.050

Van Waes, L., Leijten, M., Pauwaert, T., & Van Horenbeeck, E. (in press). A Multilingual Copy Task: Measuring Typing and Motor Skills in Writing with Inputlog. *Journal of Open Research Software*.

Van Waes, L., & Schellens, P. J. (2003). Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers. *Journal of Pragmatics*, *35*(6), 829–853.

Wengelin, A. (2007). The word-level focus in text production by adults with reading and writing difficulties. *Studies in Writing*, *20*, 67. Retrieved from https://doi.org/10.1163/9781849508223_006

Wobbrock, J. O., & Myers, B. A. (2006). Analyzing the input stream for character-level errors in unconstrained text entry evaluations. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *13*(4), 458–489. Retrieved from https://doi.org/10.1145/1188816.1188819

Xu, C. (2018). Understanding online revisions in L2 writing: A computer keystroke-log perspective. *System*, *78*, 104–114.

Zhu, M., Zhang, M., & Deane, P. (2019). *Analysis of keystroke sequences in writing logs* (ETS Research Report 19-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12247