The Keys to Writing

Conijn, Rianne

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

The Keys to Writing

A writing analytics approach to studying writing processes using keystroke logging

Rianne Conijn

The Keys to Writing

A writing analytics approach to studying writing processes using keystroke logging

Rianne Conijn

The Keys to Writing
A writing analytics approach to studying writing processes using keystroke logging

Rianne Conijn
PhD Thesis
Tilburg University and University of Antwerp 2020

Cover illustration & design: Jordi Bombeeck
Print: Ridderprint | www.ridderprint.nl
ISBN: 978-94-6416-083-3

SIKS Dissertation Series No. 2020-23
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

The Keys to Writing

A writing analytics approach to studying writing processes using keystroke logging

De Sleutel tot Schrijven

De studie van schrijfprocessen met behulp van toetsaanslaganalyse

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus, prof.dr. K. Sijtsma
en Universiteit Antwerpen
op gezag van de rector magnificus, prof.dr. H. Van Goethem,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie
in de Aula van Tilburg University
op vrijdag 16 oktober 2020 om 13.30 uur

door

Maria Anna Conijn

geboren op 17 januari 1992 te Alkmaar.

Promotores

Prof.dr.ir. Pieter Spronck, Tilburg University

Prof.dr. Luuk Van Waes, University of Antwerp

Prof.dr. Menno van Zaanen, South African Centre for Digital Language Resources

Committee

Dr. Laura Allen, University of New Hampshire

Prof.dr. Eva Lindgren, Umeå University

Prof.dr. Sven De Maeyer, University of Antwerp

Prof.dr. Mykola Pechenizkiy, Eindhoven University of Technology

Prof.dr. Marc Swerts, Tilburg University

Dr. Mark Torrance, Nottingham Trent University

Contents

1

Introduction

Writing is omnipresent in our society and plays, more than ever, an important role in our daily communication, work, and learning (Brandt, 2014). As Deborah Brandt puts it, millions of people (including myself) spend more than half of their working day "with their hands on keyboards and their minds on audiences" (Brandt, 2014, cover). However, teachers and employers often complain about the poor written communication skills of graduates (Buckingham Shum et al., 2016). In addition, several studies showed that students have difficulties with creating academic texts (e.g., Lea & Street, 1998; Mateos & Solé, 2009).

Insight into writing processes, or the cognitive and behavioral actions involved in writing, allows for a better understanding of the difficulties students face during writing. For example, such insight could indicate when, where, and why writers struggle (see e.g., Likens et al., 2017). This knowledge could in turn be used for feedback and instruction on the writing process. Feedback and instruction on the writing process, as opposed to the writing product, is important for three main reasons. First, instruction on the writing process frequently results in higher writing quality, compared to other types of instruction (Graham & Perin, 2007). Second, insight into the writing process can enhance students' awareness of their writing progress, and thereby improve effective development of task strategies (Hattie & Timperley, 2007) as well as students' ability to self-regulate their writing (Fidalgo & Torrance, 2017). Finally, as the feedback and instruction is not aimed at a single writing product, the developed strategies and skills could be more easily applied to other tasks (Schunk & Swartz, 1993).

Unfortunately, it is often difficult or even impossible for teachers to gain access to students' writing process, especially in large classrooms or online settings. Keystroke logging has been increasingly used as a scalable and unobtrustive solution for this. With keystroke logging, every key pressed on a keyboard during writing is recorded, resulting in a detailed and timed overview of each key typed by a student (Leijten & Van Waes, 2013; Lindgren & Sullivan, 2019). The analysis of these keystroke logs, keystroke analysis, can provide insight into students' writing processes.

However, there is a large gap between these fine-grained keystrokes and the higher-level writing processes as proposed in the process models of writing (Galbraith & Baaijen, 2019). In addition, it is still largely unknown how these detailed log data could be used to provide teachers with meaningful insight into the writing process. Therefore, the current

dissertation aims to identify how keystroke logging can be used to gain meaningful insight into students' writing processes. Here, I specifically focus on higher education students, hereafter referred to as students, and higher education lecturers, hereafter referred to as teachers.

In this dissertation, I address this aim using a writing analytics approach. Writing analytics is defined as "the measurement and analysis of written texts for the purpose of understanding writing processes and products, in their educational contexts" (Buckingham Shum et al., 2016, p. 481). The field of writing analytics can be considered as a subfield of the more developed fields of learning analytics and educational data mining, which analyze data about learners and their context, to improve learning and teaching in general (Clow, 2013; Romero & Ventura, 2013). Like the advocated approach in these fields, this thesis is data-driven, but grounded in (writing process) theories, to motivate methodological choices and to enhance the interpretation of the findings (Gašević et al., 2015). Accordingly, in this introduction, I provide an overview on writing process theories, followed by a review of existing literature on the measurement and analysis of writing processes, specifically using keystroke logging. Finally, I provide an overview of this dissertation.

1.1 Writing process theories

The first—and probably most widely used—model on writing processes is the model proposed by Flower & Hayes in 1980. This model consists of three cognitive writing processes: planning, translating, and reviewing, which are influenced by the long-term memory and the task environment. Planning includes the generation of ideas, organization, and goal setting; translating describes the process of translating these ideas into (written or typed) language; and reviewing includes the evaluation and revision of the text produced so far. After this initial model, several refinements were made and alternative models have been proposed (for a detailed review see Alamargot & Chanquoy, 2001; Becker, 2006).

Two major refinements of Flower & Hayes' (1980) model are Hayes' (1996) model and Hayes' (2012) model. Hayes (1996) distinguished two components of the writing process: (1) the task environment, consisting of the social and physical environment; and (2) the individual, consisting of the working memory, motivation and affect, and cognitive processes. Working memory was added as a central part in writing. In addition, motivation and affect are included as factors influencing writing. Lastly, the cognitive processes are redefined into text interpretation, reflection, and text production. In Hayes' (2012) model,

three different levels were distinguished: (1) the control level, which consists of motivation, goal setting, plans, and writing schemas; (2) the process level, divided into the writing processes: proposer, evaluator, translator, and transcriber; and the task environment, consisting of collaborators and critics, transcribing technology, task materials, written plans, and text-written-so-far; and (3) the resource level, which consists of attention, working memory, reading, and long-term memory. The two major changes to Hayes' (1996) model are the addition of the transcriber and the removal of the planning and reviewing process. Hayes (2012) included the transcriber, or the process of putting the ideas (translated into words and sentences) on paper, as transcription competes for cognitive resources as well; and moreover, the transcription mode (e.g., handwriting or type of keyboard) plays an important role in the writers' environment. The planning and revision processes were removed as he sees these as specialized writing activities, which also consist of the processes of proposing, evaluating, translating, and transcribing. In 2014, this model has been extended to encompass for visual components in writing, by adding the visual design schemas at the control level (Leijten et al., 2014). In addition, the searcher was added as process in the writing process level, to denote the writers' searching process for information. Lastly, motivation management was added in the resource level, to account for tasks over extended periods of time. This latest model of writing processes is shown in Figure 1.1.

In addition to the refinements of the model describing the full writing process, specific models were created for certain subprocesses in writing. Flower et al. (1986) developed a model specifically on revision processes, in which the writer starts with the task definition, a plan on how to guide revision. Then the writer reads the text written so far, to comprehend and evaluate whether their writing goals are met. This results in a problem representation which can be ill-defined, merely a detection of the problem, or well-defined, a diagnosis of the problem. Based on the problem representation, the problem can either be ignored, or a strategy will be selected to solve the problem: rewrite or revise (Flower et al., 1986).

Moreover, several alternative writing models were proposed. Here, I only describe three of the most influential models. First, Bereiter and Scardamalia (1983; 1987) developed a model focused on the development of writing. Their model distinguishes two writing strategies, which can bee seen as two extremes on a continuum: the knowledge telling strategy and the knowledge transforming strategy. The knowledge telling strategy is a strategy used by novices, where ideas are formulated without reorganization of content

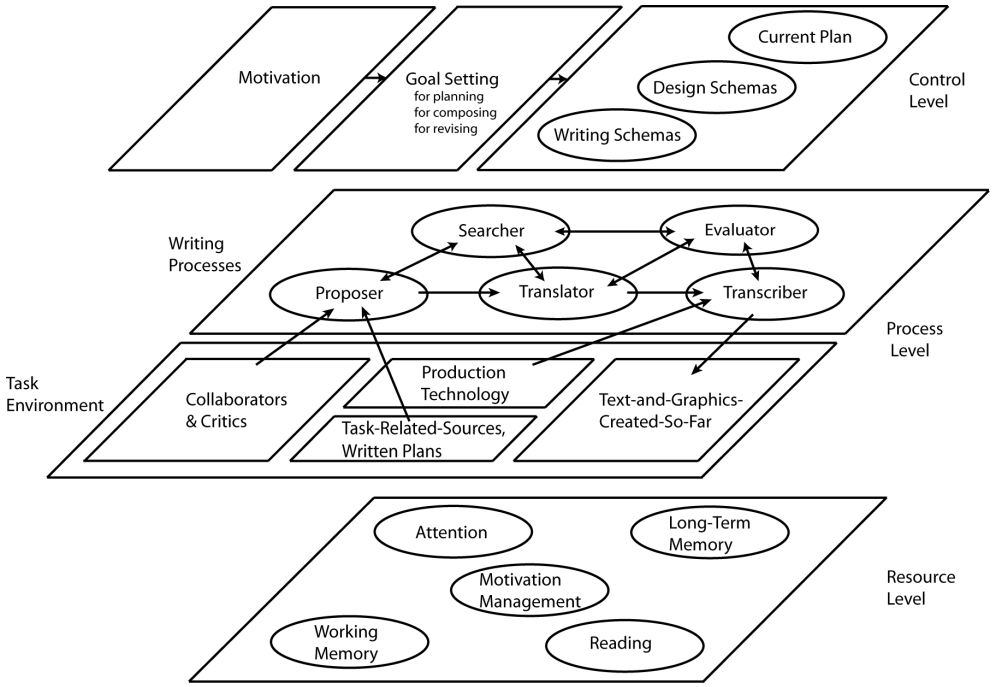

Figure 1.1: Latest writing process model (adapted from Leijten et al., 2014, with permission of the rights holder, M. Leijten).

or linguistic features. The knowledge transforming strategy is a strategy used by experts, where content and linguistic features are continuously re-elaborated until they are deemed appropriate according to the author's intentions and goals. Later on, this model was extended with a third strategy: knowledge crafting. This strategy is employed by even more expert writers, where the re-elaboration is not only focused on the author's intentions or goals, but also on the imagined reader's interpretation of the text (Kellogg, 2008). Second, Kellogg's (1996) model emphasizes the influence of the working memory on writing and details how it supports cognitive writing processes. He distinguishes three cognitive processes: the formulation of ideas, the execution of motor processes (e.g., typing, handwriting), and the monitoring of these processes. Lastly, Van Lier (2000) stresses that the cognitive processes need to be explored within the context of the writers' environment, as well as their (social) relationships and interactions within this environment. This can be seen as a socio-cognitive model of writing.

All these models show that writing involves a variety of processes, ranging from low-level (peripheral) processes, e.g., motor processes such as typing, to high-level (central) processes, e.g., text evaluation (Olive, 2014). However, the models do not specify how these processes are coordinated or timed. Given the wide variety of processes that need to be coordinated, writing can be cognitively highly demanding (Olive & Kellogg, 2002). It is generally assumed that writing processes can happen concurrently when the demands do not exceed the available cognitive resources (Hayes, 2012; Olive, 2014). Olive (2014) further specifies the timing and coordination of these processes using the parallel and cascading model of writing. Within this model, it is argued that although text is produced incrementally, different segments are produced in parallel (when sufficient cognitive resources are available), where information flows from high-level to lower-level processes (Olive, 2014). For example, one segment can be written down, while the next segment is being formulated. Efficient coordination of the writing processes is important for the quality of the writing and can be improved by minimizing the concurrent demands (Olive, 2014; Torrance & Galbraith, 2006). This can, for example, be done by improving (automating) lower-level skills (e.g., keyboarding), enhancing memory retrieval skills, and by using writing strategies to divide writing into subtasks (e.g., note taking; Torrance & Galbraith, 2006).

To conclude, writing is a complex process, including a wide variety of subprocesses that have to be coordinated within the limited pool of cognitive resources available. Accordingly, there is a considerable amount of literature on how this process and the different subprocesses can be measured and analyzed.

1.2 Measuring writing processes

Writing processes can be measured in two ways: synchronously, during the writing process, also known as online measurement, and asynchronously, after the writing process (Janssen et al., 1996; Leijten & Van Waes, 2013). Commonly, self-report methods such as thinking-aloud (synchronous) and retrospective interviews (asynchronous) have been used (see e.g., Plakans, 2009; Solé et al., 2013). These methods provide (relatively) direct evidence on writing processes, however, they have been criticized for being intrusive. Therefore, other, less-intrusive measures have been used as well, such as observations or screenlogs (synchronous, see e.g., Xu & Ding, 2014). These are more indirect measures, as they do not directly measure writing processes, but inferences can be made about the cognitive processes involved. Although all these methods have proven to be effective measures of the writing process,

they are time-intensive, and hence non-scalable. Accordingly, they cannot easily be used in the classroom to determine students' writing strategies.

Nowadays, real-time data can be collected automatically during writing, making it possible to collect information on the writing process in an unobtrusive and scalable way. Examples of information that can be automatically extracted, are keystroke presses (keystroke logging), mouse clicks (clickstream logging), and eye movements and fixations (eye tracking). These are also more indirect measures, hence again inferences need to be made about the underlying processes. In this thesis, I focus on the use of keystroke logging to analyze writing processes. Specifically, I investigate the keystrokes made by typing on a computer or laptop keyboard, and do not include other input modes, such as handwriting, smartphone keyboards or touch interfaces.

1.3 Collecting keystroke data

To date, there are multiple stand-alone and web-based programs that can log keystroke data. These programs log the specific key (e.g., 'Alt', 'k', or '$', sometimes in the form of a key code), the key press time, and the key release time (in milliseconds) for every key pressed. This results in a sequence of timestamped keystroke data: a keystroke log. Examples of keystroke logging tools are Trace-it, ScriptLog, Inputlog, CyWrite, and EyeWrite (Van Waes et al., 2012). In addition to keystrokes, some of these tools allow for the collection of additional data, such as the force applied when pressing the keys, mouse movements, eye movements, speech, and use of digital sources (e.g., websites, dictionaries, and other documents). Furthermore, some tools provide built-in analyses and replays of the text production (Van Waes et al., 2012).

Keystroke log data provide objective, real-time, and fine-grained information on writers' unfolding typing process during text composition. As keystroke data can be collected automatically, this metric is more scalable and less intrusive than traditional thinking-aloud methods and observation studies. However, as keystroke data are more indirect measures of the writing process, it less directly interpretable, compared to more direct measurements such as thinking-aloud (Galbraith & Baaijen, 2019). Therefore, the correct analysis of keystroke data is of key importance.

1.4 Analyzing keystroke data

Keystroke data have been analyzed for a wide range of objectives, including writer identification and authentication (Karnan et al., 2011), prediction of performance in programming tasks (Thomas et al., 2005), prediction of writing quality or essay scores (M. Zhang et al., 2016), prediction of task complexity (Grabowski, 2008), detection of emotional states (Bixler & D'Mello, 2013; Salmeron-Majadas et al., 2014), detection of deceptive writing (Banerjee et al., 2014), analysis of writing fluency (Abdel Latif, 2009; Van Waes & Leijten, 2015), diagnosing Alzheimer's disease (Van Waes et al., 2017), and relating the writing process to the linguistic features in the writing product (Allen, Jacovina, et al., 2016). Moreover, several studies have shown that keystroke data can indeed be used for real-time information on the writing process (e.g., Baaijen et al., 2012; Tillema et al., 2011; Van Waes et al., 2014).

Given the fine-grained nature of the keystroke data, feature extraction (i.e., variable extraction) is necessary before the keystroke data can be analyzed. The features extracted in previous work can be broadly organized into five categories: (1) features related to pause timings or latencies, such as interkeystroke intervals (IKI) between or within words (see e.g., Medimorec & Risko, 2017) or initial pause time (see e.g., Allen, Jacovina, et al., 2016); (2) features related to revising behavior, such as the number of backspaces (see e.g., Deane, 2014); and (3) features related to fluency or written language bursts; i.e., sequences of text production without interruptions, such as the number of words per burst after a pause or revision (see e.g., Baaijen et al., 2012; Van Waes & Leijten, 2015); (4) features related to verbosity, such as the number of words (see e.g., Allen, Jacovina, et al., 2016); and (5) features related to other events, such as digital source usage (see e.g., Leijten, Van Waes, et al., 2019).

The feature selection and following analysis of the features highly depends on the methodological approach. These approaches can be roughly divided into theory-driven and data-driven approaches. Theory-driven studies use a select set of keystroke features, often triangulated with other data such as manual annotations or thinking-aloud, within a tightly controlled experimental setting. For example, several studies link keystroke data to the three writing processes as defined by Flower & Hayes (1980): planning, translating, and reviewing processes (see e.g., Tillema et al., 2011). Likewise, studies have used a dual task approach, in which they measured the performance on a second task, which had to

be performed simultaneously with the writing task, to determine the cognitive demands in writing (see e.g., Alves et al., 2007). The down-side of these theory-driven studies is that they are typically time-intensive and do not allow for scalable systems.

By contrast, data-driven studies often include as many features as necessary to build an accurate model for the problem at hand. Studies using keystroke analysis for authentication and identification can be considered examples of these data-driven approaches (see e.g., Bixler & D'Mello, 2013; Karnan et al., 2011). Although this data-driven approach can result in highly accurate automatic detection systems, this does not always provide insight into the model underlying the prediction, and hence does not provide insight into the phenomenon under study.

Accordingly, the current dissertation uses a data-driven approach, informed by writing process theory as well as stakeholders' needs to identify which features need to be selected from the keystroke data, as well as to interpret the results. With this approach I do not aim to make any theoretical claims, nor do I directly link the primarily behavioral keystroke data with the cognitive writing processes. Rather, I focus on an automated, and scalable solution to provide stakeholders with insight into writing processes, without limiting the research to a specific genre or language.

1.5 Structure of this dissertation

This dissertation aims to answer the main research question:

> **Main research question:** *How can keystroke logging be used to gain meaningful insight into students' writing processes?*

This main research question is addressed by four subquestions, which are answered with six studies divided over six chapters.

1.5.1 Identifying stakeholders' needs

Since there are many subprocesses of writing that can be active concurrently, we need to determine which insights into the writing process stakeholders desire:

> **Subquestion 1:** *What indicators of students' writing processes are considered desirable, according to multiple stakeholders, for providing feedback on the writing process?*

In **Chapter 2** we investigate the indicators of students' writing processes that are perceived as desirable for the design of systems that provide automated, personalized feedback. In addition, we provide use cases of how this feedback can be integrated into teaching and learning practices. To elicit these indicators and use cases, participatory consultation sessions were conducted with five representative groups of stakeholders: bachelor students, PhD students, teachers, writing specialists, and professional development staff.

1.5.2 Determining capabilities of keystroke analysis

Next, we need to determine what is technically feasible, given the keystroke data available. This is addressed in chapters 3 and 4, with the following research question:

> **Subquestion 2:** *What keystroke features can be used to gain insight into students' writing processes?*

Chapter 3 determines the sensitivity of frequently used keystroke features across tasks with different cognitive demands. Bayesian linear mixed effects models are used to determine the differences in keystroke features between two tasks in two datasets: one consisting of a copy task and an email writing task, and one with a larger difference in cognitive demand: a copy task and an academic summary task. This provides insight into which keystroke features would be of interest for gaining insight into students' (cognitive) writing processes.

In **Chapter 4** we identify which keystroke features can be used to predict writing quality. Specifically, we determine whether keystroke logging can be used to identify students at risk already during the writing process. Machine learning models are trained to predict writing quality. In addition, we identify which features are important for the early prediction, and how this feature importance changes during the writing process.

1.5.3 Gaining insights

While considering the stakeholders' needs and the possibilities and limitations of the keystroke data, we then turn to modeling students' writing processes. Here, we scope the dissertation to specifically focus on revision processes only, as this one of the most desired processes to gain insight into by stakeholders, which are rather directly observable in the keystroke data. Accordingly, chapters 5 and 6 address the following research question:

Subquestion 3: *How can we model keystroke features to gain insight into students' revision processes?*

Chapter 5 provides a comprehensive product-oriented and process-oriented tagset of revisions in writing. Current advances in data collection and analysis, such as keystroke logging, eye tracking, and natural language processing, have made it possible to gain a more complete and in-depth analysis of revision. Yet, a complete overview of and approach to extracting all these features is lacking. Therefore, this chapter reviews the revision taxonomies used in writing studies and summarizes them in ten categories of revisions. In addition, to make these categories measurable, we describe how both manual annotation and automatic extraction can be used to collect features related to these categories.

In **Chapter 6** we automatically classify one of the smallest types of revisions from the revision taxonomy: typographic error revisions. On the one hand, these types of revisions are low-level, and hence less important, so we would like to be able to ignore them in the analysis. On the other hand, typographic errors, and especially the revision of these 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. This chapter uses machine learning to model these typographic error revisions.

1.5.4 Operationalizing insights

With the insights obtained from modeling the keystroke data we return to the stakeholders, to determine how these models need to be presented and integrated into the learning design, to ultimately improve the learning and teaching of writing:

Subquestion 4: *How can we visualize students' revision processes in order to make them actionable for teachers?*

In **Chapter 7** we run several co-creation sessions with writing teachers in order to determine how to visualize students' writing processes, and specifically the revision process. In addition, we provide guidelines on how these visualizations should be integrated into teaching and learning practices.

1.5.5 Discussion

Finally, in **Chapter 8** I conclude with a general discussion on the findings of all the studies presented in this dissertation. In addition, overarching implications and reflections are provided on the use of keystroke logging in writing research and writing education, as well as opportunities for future work.

1.6 Academic integrity

The research on which this dissertation is based and the dissertation itself complies with the standards for good research practices as defined in the current Netherlands Code of Conduct for Research Integrity (2018). In this section I further specify some of the choices I have made in this respect.

1.6.1 Authorship statement

The first and last chapter of this dissertation (Introduction and Discussion), are solely written by me, and hence the personal pronoun 'I' is used in these chapters to refer to me, the author. All other chapters are based upon co-authored articles (published or under review), and hence within these chapters (and within references to these chapters) the personal pronoun 'we' is used to refer to the authors.

For all the co-authored chapters, I am the first and main author, and have been responsible for the full process, including conceptualization, methodology, data collection, data pre-processing, data analysis, and writing. Luuk Van Waes and Menno van Zaanen, as my promoters, have aided me in the conceptualization and have reviewed my writing. For the wide range of methodologies and analyses used within this dissertation, I have sought advice from experts in the field. These experts and hence co-authors have supported me with methodology (participatory sessions Chapter 2: Roberto Martinez-Maldonado, Simon Knight, and Simon Buckingham Shum), data collection and annotation (Chapter 4: Chrissy Cook; Chapter 5: Emily Dux Speltz and Evgeny Chukharev-Hudilainen; Chapter 6: Mariëlle Leijten), data pre-processing (Chapter 5: Evgeny Chukharev-Hudilainen), and data analysis (Bayesian linear mixed effects models Chapter 3: Jens Roeser). All co-authors have reviewed my writing.

1.6.2 Ethics in data collection and data sharing

Within my dissertation, I have collected various types of data, including audio recordings of focus groups and interviews, questionnaire data, and keystroke data. In Chapter 3, the data have been collected from an anonymized fully open dataset, and for Chapters 5 and 6, I have used datasets made available by my colleagues wihtin international collaborations. For Chapters 2, 3, 4, and 7, I have collected new data. The studies in these chapters have been approved by the school-level Research Ethics and Data Management Committee. All participants provided informed consent before participation, and were debriefed after participation. All data were anonymously collected and stored.

Given the highly sensitive nature of keystroke data, some additional precautions were taken to ensure participants' privacy. Participants were required to provide a second consent after finishing the writing task in which their keystrokes were logged. Within this consent, they needed to indicate that they did not type any personal information during the task, and they still agreed with their data being used. If not, their data were destroyed. However, as it could not be guaranteed that no sensitive data might have been typed during the task, the keystroke data were not made openly available nor shared outside the research team. Only the highly anonymized and aggregated revision dataset, including the type of revision and timing of the revision (but excluding the actual content of the revision), was made openly available.

2

Desired indicators to provide feedback on the writing process

Adapted from: Conijn, R., Martinez-Maldonado, R., Knight, S., Buckingham Shum, S., Van Waes, L., & van Zaanen, M. (under review). *How to provide automatic feedback on the writing process? A participatory approach to writing analytics design.*

Feedback on students' writing is an emerging theme in developing writing tools. However, writing support tools tend to focus on assessing final or partial, intermediate products, rather than the writing process. Keystroke logging can enable provision of automated feedback during, and on aspects of, the writing process. Despite this potential, little is known about the critical indicators for providing this feedback. Therefore, this chapter proposes a participatory approach, to identify the indicators of students' writing processes that are meaningful for educational stakeholders, and that can be included in the design of systems that provide automated, personalized feedback in higher education. This approach is illustrated through a qualitative research design that included five participatory sessions with five distinct groups of stakeholders: bachelor and postgraduate students, teachers, writing specialists, and professional development staff. Results illustrate the value of the proposed approach, showing that students are especially interested in lower-level behavioral indicators, while other stakeholders focus on higher-order cognitive and pedagogical constructs. These findings lay the groundwork for future work in extracting these higher-level indicators from in-depth analysis of writing processes. In addition, key stakeholder differences in terminology used and the levels at which the indicators are discussed, highlighting the need for human-centered, participatory approaches to design and develop writing analytics tools.

Acknowledgements. The study presented in this chapter was partially funded by the Australia Awards Endeavour Research Fellowship (grant number 6381-2018).

2.1 Introduction

Academic writing plays a critical role in higher education, but it is a difficult skill for students to develop (Ferris, 2011; Staples et al., 2016). Several meta-analyses have shown that strategy instruction is one of the most effective interventions in improving writing (Graham & Perin, 2007; Graham et al., 2012), where strategy instruction is defined as "explicitly and systematically teaching students strategies for planning, revising, and/or editing text" (Graham & Perin, 2007, p. 449). For strategy instruction, and especially for strategy instruction aimed at writers in higher education who already adopted some (un)successful writing strategies, it is important to gain insight into students' writing processes; the cognitive and behavioral actions involved in writing. This allows teachers to comment on students' writing strategies, to let students reflect on the current strategies, and to teach new and more effective strategies.

However, it is often difficult or even impossible for teachers to gain access to students' writing processes, especially in large classrooms or online settings. That is probably one of the main reasons that, up till now, most teachers focus their feedback on text or product characteristics. Moreover, the amount of writing studies that focus on the relation between text characteristics and text quality outnumbers by far the studies on writing processes (cf. Crossley, 2020). Likewise, it is difficult for students to gain insight into their own writing processes, as some processes might be implicit and not reach students' awareness. Some insight into these processes can be gained via direct observations, video analysis, or think-aloud protocols (e.g., Solé et al., 2013; Tillema et al., 2011). However, these approaches are time-intensive and not scalable.

Therefore, automated data collection, such as keystroke logging, has been increasingly used to shed light on the processes writers follow to create their final writing products. Keystroke analysis has been proposed as a scalable solution to help teachers gain insight into students' writing processes (Ranalli et al., 2018a). The analysis of the timing of every key pressed during the writing process has been used to identify higher-order writing strategies, for example, to identify usage of external sources (Leijten & Van Waes, 2013); planning and reviewing behavior (Deane, 2014; Medimorec & Risko, 2017); and written language bursts (Baaijen et al., 2012). Thus, keystroke data can be used, at least to some extent, to build intelligent, computer-based systems that are designed to support writing either during or at the end of the writing process. However, the current indicators extracted

from keystroke data are still relatively low-level behavioral features, such as keystroke frequencies or timings between keystroke events. These indicators require other sources of contextual information to be meaningful or to point at critical cognitive processes (Galbraith & Baaijen, 2019).

Therefore, we need to get better insight into which indicators need to be extracted from keystroke or alternative sources of data to gain valuable insight into students' writing processes for informing strategy instruction. For this, we argue that it is important to first determine what indicators of the writing process are desired by different stakeholders (e.g., teachers and students) according to their learning or pedagogical goals. These indicators in turn can be assessed to identify whether they are useful and technically feasible to be obtained. No prior studies have systematically examined which indicators of the writing process can be useful to support teaching and learning, according to stakeholders' needs. Therefore, the current chapter proposes a participatory approach to identifying what evidence would be useful to extract from the writing process and its potential instructional uses in higher education. These indicators can ultimately be used to develop a computer-based system designed to support writing, a writing analytics tool (or writing tool in short).

Automated, personalized writing tools are hard to develop for writing, for two reasons. First, as writing is an ill-defined domain (Allen et al., 2015; Steenbergen-Hu & Cooper, 2014), writing specialists need to be involved in the development of such tools (Cotos, 2015). Second, these systems are used less and are less effective if they are not integrated into instructors' learning design (Link et al., 2014). Most of current studies reporting on writing tools have included specialists, teachers, and other stakeholders only after the development of such tools (El Ebyary & Windeatt, 2010; Rapp & Kauf, 2018; Roscoe et al., 2014). By contrast, in this chapter we present a study which illustrates our participatory approach by conducting participatory sessions with educational stakeholders before the design of writing analytics tools. This chapter aims to determine what indicators of students' writing processes are desirable to provide automated, personalized writing feedback in higher education and how these can be connected with teachers' learning designs.

2.1.1 Writing process models

Writing processes have been deeply studied over the past decades (Bereiter & Scardamalia, 1987; Flower & Hayes, 1981; Hayes, 1996, 2012; Kellogg, 1996), for an overview, see Becker

(2006). In this chapter, we adopt Flower & Hayes' (1981) model, as this is the most pragmatic model for our study. This model distinguishes three different writing processes: planning, translating, and reviewing. Planning consists of the generation of ideas, organization, and goal setting; translating describes the process of translating these ideas into (written or typed) language; and reviewing consists of evaluating and revising the text produced so far.

These cognitive processes are highly dependent on the writers' environment, and hence need to be explored within the context of this environment (Van Lier, 2000). In addition, these cognitive processes are not randomly distributed over the time of the writing process, and hence need to be explored in relation to time, or when they occur during the writing process (Rijlaarsdam & Van den Bergh, 1996). Specifically, time needs to be considered because it might give more information about the purpose of the process, and sequences of cognitive processes differ across writers (Rijlaarsdam & Van den Bergh, 1996). Therefore, we also discuss the cognitive processes in relation to time and the aspects of the writers' task environment, as described in Hayes (2012): collaborators and critics; transcribing technology; task materials and written plans; and the text written so far.

2.1.2 Keystroke data

Keystroke analysis has been shown as a useful tool to gain insight into the writing process (Leijten & Van Waes, 2013; Lindgren & Sullivan, 2019). However, keystroke data have been criticized because it is hard to associate the low-level behavioral actions with higher-level cognitive processes (Galbraith & Baaijen, 2019). Yet, various elements, such as pauses, revisions, and production bursts have been related to theory and models on writing processes.

Pauses, and in particular pauses between words and between sentences (rather than within words), have been related to Flower & Hayes' (1980) planning and reviewing processes (Baaijen et al., 2012; Medimorec & Risko, 2017). Here, longer pauses are suggested to indicate a higher cognitive effort (Van Waes et al., 2014; Wallot & Grabowski, 2013; Wengelin, 2006). To describe these pauses with keystroke data, often several summary statistics are computed (e.g., mean, standard deviation, maximum) of interkeystroke intervals (IKI), the time from a key press until the next key press. Revisions have been related to reviewing processes (Van Waes et al., 2014) and linearity of the writing process (Baaijen & Galbraith, 2018). To describe these revisions, counts, durations, and ratios of (sequences of)

backspace and delete keys are often used, and percentages of keystrokes typed at the leading edge (Baaijen & Galbraith, 2018; Deane, 2014). Lastly, bursts are described as part of Flower & Hayes' (1980) translation processes; sentences are composed in sentence parts or bursts, sequences of text production without a long pause (Kaufer et al., 1986). Longer and more bursts have been related to higher writing proficiency (Deane, 2014).

Thus, keystroke data can be used, at least to some extent, to automatically gain insight into writing processes. However, the current variables extracted from keystroke data are still relatively basic frequency and timing variables, which may not be directly useful to improve writing feedback and writing instruction. Therefore, in this study we identify what elements of students' writing processes are desirable for providing feedback on the writing process. Ultimately, these indicators could be used to develop writing analytics tools.

2.1.3 Writing tools

Providing personalized and timely feedback on writing is a time-intensive task for teachers. To address this problem, a wide variety computer-based systems have been developed to support writing instruction and assessment (for an overview, see Allen et al., 2015). Three main categories of writing tools have been identified based on their functionality: automated essay scoring (AES), automated writing evaluation (AWE), and intelligent tutoring systems (ITS; Allen et al., 2015). AESs are grading systems that can be used for summative assessment, to replace or assist teachers in assessing writing quality (Dikli, 2006), for example *e-rater* (Attali & Burstein, 2006). In comparison, AWEs are intended as formative assessment tools, providing more detailed feedback and correction suggestions (Cotos, 2015), for example *Criterion* (Link et al., 2014) and *AWA* (Knight et al., 2017). ITSs are the most complex systems, providing not only feedback, but also include instructional elements, interactivity, and probing questions (Ma et al., 2014). ITSs are widely available in domains such as mathematics and business, but less in more ill-defined domains such as reading and writing (Steenbergen-Hu & Cooper, 2014). Two examples of ITSs targeted at supporting writing are *eWritingPal* (Roscoe et al., 2014) and *ThesisWriter* (Rapp & Kauf, 2018).

All three types of systems have been extensively studied in the writing context, and have been shown to enhance student motivation, autonomy, and improve writing quality (Cotos, 2015). However, the majority of these systems use a product-oriented approach, in

which feedback is provided on students' written products (Cotos, 2015; Wang et al., 2013). Some tools do provide additional resources to aid the writing process. For example, Criterion provides a portfolio history of drafts, to have insight into one's writing progress over time (Link et al., 2014); eWritingPal includes lecture videos with animated agents to teach strategies for pre-writing, drafting, and revising (Roscoe et al., 2014); and ThesisWriter uses scaffolding to provide instructions on strategies for research report writing (Rapp & Kauf, 2018). However, these tools do not yet collect evidence from writing process nor provide feedback on specific writing processes.

In addition, these tools are usually only evaluated after the development of the tool (see e.g., El Ebyary & Windeatt, 2010; Rapp & Kauf, 2018; Roscoe et al., 2014). However, it has been argued that it is not enough to introduce stakeholders after the development; stakeholders need to be included early on in the design process (Dollinger et al., 2019). By including information from writing specialists to identify why and how particular affordances are needed, rather than simply including all features that are technically feasible, the design could be improved (Cotos, 2015). In this way, the design can also be better tuned to the educational context (Conde & Hernández-García, 2015). When writing tools are tuned to the educational context, they are perceived more positively by students, resulting in a higher adoption (Shibani et al., 2019). Therefore, there has been a growing interest in including the voices of educational stakeholders early on in the design of writing analytics tools, and learning analytics tools in general (e.g., Buckingham Shum et al., 2019; Martinez-Maldonado et al., 2015; Wise & Jung, 2019).

2.1.4 Current approach

In the current chapter, a qualitative research design is implemented to identify what evidence would be useful to extract from the writing process and its potential instructional uses in higher education. Recently, the importance of qualitative research in computer assisted language learning has been stressed, as it can inform the design, development, and evaluation of language tools through a deeper understanding of the stakeholders involved (M. Levy & Moore, 2018). As a result, participatory sessions using the focus group technique (Kidd & Parshall, 2000) are suggested to be conducted to gain insight into stakeholders' perspectives before the writing tools are designed. Within a participatory approach it is important to include different groups of stakeholders, as the ideas might differ across stakeholders (Woolner et al., 2007). Different stakeholders in writing instruction have shown

to feature quite different perceptions on academic writing (Itua et al., 2014; Lea & Street, 1998; Wolsey et al., 2012). For example, students have indicated content and knowledge as the two most important criteria items for assessing essay writing (Norton, 1990), while teachers consider argument and structure to be the key items they use in their assessments (Lea & Street, 1998; Norton, 1990).

The proposed approach is illustrated through a study with five groups of stakeholders who would consult automated reports on students' writing: bachelor students, PhD students, teachers, professional development staff, and writing researchers. Bachelor and PhD students were chosen, to represent groups of students with relatively low and relatively high experience in academic writing, respectively. More expert writers tend to be more strategic in their writing processes, compared to novice writers (Kaufer et al., 1986), and hence might desire insight in different types of indicators of their writing process. Teachers and professional development staff were included, to identify desired indicators from the teacher and teacher trainers' perspective. Lastly, writing researchers were included, to identify desired indicators from writing research and theory, and to better connect writing analytics to educational practice (cf. Buckingham Shum et al., 2016). Outcomes of the sessions are mapped to (1) writing tool development, to inform the potential further design of one or more writing tools; and to (2) keystroke data, to inform the use of keystroke data in education and writing research. This illustrates how a human-centered approach can be adopted into the particular context of writing, which can also be useful for the broader area of learning analytics.

2.2 Method

2.2.1 Participants

Five participatory sessions were conducted with five representative groups of stakeholders. In total 25 stakeholders participated: 4 university teachers, 5 bachelor students, 6 PhD students, 6 professional development staff (teacher trainers) and 4 writing experts (writing researchers). The bachelor students were recruited via the university's participant pool. The teachers and professional development staff were recruited by email via the university's language center. The PhD students and writing researchers were recruited via the authors' personal network. Both bachelor and PhD students were selected based on whether they completed an academic writing course. Teachers were selected based on their years of ex-

Table 2.1: Example of the use case provided to the participants

Question	Example
Context	I have to complete a writing assignment within a specified word limit
When	I am working on the assignment, and I exceed the word limit
What	an automatic tool within the word processing software
Who (Addresses)	me
How (By)	providing a pop-up stating that I exceeded the word limit and have to cut some words before submitting the assignment
Why (Outcome)	to make sure I will not submit a writing assignment which is too long.

perience (> 10 years) in teaching academic writing, professional development staff were selected based on their years of experience in teacher training (> 5 years), and writing experts were selected based on their years of experience in writing research (> 2 years). Students came from the fields of Sociology, Communication, Cognitive Science, and Artificial Intelligence. Teachers and professional development staff worked across a wide variety of fields, including Arts, Social Sciences, Business, Law, Science, and Engineering, teaching both first and second language learners.

2.2.2 Materials and procedure

After the participants provided informed consent, participants were asked to fill out a short demographics' questionnaire. Thereafter, the goals, procedure, and rules for the focus group were explained. The focus group consisted of two parts, focused on the respective research questions. For these two parts, a semi-structured, open-ended schedule was developed.

The first part focused on capturing participants' perspectives on the writing process and how evidence about the writing process could be used to support teaching and learning. In the sessions with teachers, writing researchers, and professional development staff two questions were asked in the following order:

1. What do you think an instructor would like to learn about students' writing processes?
2. What do think would be useful to show a student about their writing process?

For both student focus group sessions there were three questions:

1. What would you, as a student, like to learn about your writing process?

2. What do you think an instructor would like to learn about students' writing processes?
3. What do you think an instructor should **not** see about students' writing processes?

To avoid social pressure, participants were first asked to write down their ideas on sticky notes (one idea per note). The participants got two minutes per question. Thereafter, they were asked to read their ideas out loud and discuss them (ten minutes). Participants were encouraged to write down new ideas if needed and they were asked to cluster the sticky notes with similar ideas, and to name these clusters. Lastly, participants were asked to vote for what they considered were the three best ideas.

In the second part, participants were asked to write a use case of an intervention using one or more of the ideas generated earlier. An exemplar use case was first shown for them to understand what they were supposed to generate (see Table 2.1). Then the participants had ten minutes to write their own use cases, emphasizing the context (learning design of the learning situation), state and form of the intervention (tool set, strategies/actions needed and by whom?), and expected outcomes. Afterwards, participants were given ten minutes to discuss and expand on their cases.

By the end of the session, participants had the possibility to add any further ideas or ask questions in a debrief. The sessions lasted 60–75 minutes in total. To minimize the influence of the moderators' viewpoint on the discussion, participants were encouraged to moderate the discussions themselves. When necessary, the moderator only asked open format follow-up questions, such as: *Could you provide some more details?* or *Why do you feel this is important?*

2.2.3 Analysis

NVivo 12 was used to transcribe the audio recordings of the sessions and for the qualitative analysis of the transcripts, sticky notes, and use cases (NVivo, 2015). The (clusters of) sticky notes were interpreted in the context of the dialogue. Using the coded transcripts, four of the authors analyzed which indicators of the writing process were identified, which were most important, and which were highly connected to other concepts. The importance of a topic was determined by the number of sticky notes and votes on that topic.

The codes and prevalence of the topics were compared across sessions. This was done by mapping the topics to theoretical models of writing, via thorough discussion between

the authors. This mapping was used to compare and contrast the topics of the discussions across the stakeholders. For this, we adopted one of the most widely used models of writing processes, developed by Flower & Hayes (1981), which distinguishes the three cognitive processes in writing defined above (*planning*, *translating*, and *reviewing*), as well as the '*monitoring*' process, which describes the strategic cognitive process which monitors the writer across the cognitive processes. All topics were mapped into one of the three writing processes, or into the monitoring process if the topic described monitoring or self-regulation processes associated with the three writing processes or with the writing process in general. Additionally, we indicated whether a topic was discussed in the context of an aspects of the writers' task environment, as defined by Hayes (2012): collaborators and critics; transcribing technology; task materials and written plans; and text written so far.

The topics were also categorized in terms of the level at which the indicators of the writing processes were discussed. The lowest level included behavioral indicators, which were mainly identified by the use of words related to frequency (the number of), total time spent, and occurrence of behavior (e.g., *do they plan?*). The middle level consisted of behavioral indicators that were described in the larger context of writing, for example by describing a sequence of behaviors (e.g., *how do students plan?*), behavior in relation to the writing product (e.g., *which sections required much effort?*), or behavior in relation to time or the writing process (e.g., *how do revisions change over time?*). The highest level included cognitive indicators, which were identified by the use of words such as develop, ideas, thoughts, understand, and experience (e.g., *how do ideas develop?*). Lastly, the use cases were analyzed. We compared the main focus of the use cases in each focus group, in regard to how stakeholders' ideas can be integrated into the learning design. We especially contrasted the different tool sets described, the strategies and actions needed, and the actors involved in the intervention.

2.3 Results

First, we discuss the topics of desired indicators of students' writing processes per stakeholder group individually. Next, the indicators are compared and contrasted across the stakeholder groups. Lastly, we describe the results of the use cases, which indicate stakeholders' opinions on how these indicators can be integrated into learning and teaching practices.

2.3.1 Identifying topics and ideas per stakeholder group

Bachelor students. The bachelor students wrote a total of 40 ideas on sticky notes. These were categorized into nine topics (one idea was left uncategorized because the students argued it was not related to the other ideas). An overview of the topics, ordered by the number of sticky notes, followed by the number of votes is shown in Table 2.2. Although only discussed once, typing patterns received the most votes and sticky notes of all topics. This topic was mostly related to keyboarding skills, and was the only topic that was considered desirable for both students and teachers. Planning was rated as second most important topic. The students would like teachers to know how they prepared and planned for the task and what their initial ideas were, especially to be able to receive feedback on these ideas. In addition, students would like information on the number of words and characters typed, categorized as general structure, to be able to determine whether they met the assignment requirements.

In general, students stated that teachers could use the information on students' writing process to improve instruction. For example, a student stated this as follows: "*... in terms of sentence framing, grammar usage, APA style, fonts and stuff, the teachers would want know what students' exposure is on these kinds of terms. And I think based on that, you could build a lecture or a class around it*". The students differed in opinion whether certain aspects of the writing process should remain invisible to teachers. Some stated there was nothing they wanted to hide, but others noted that they would not like the teacher to know about the grammatical errors they already fixed. Students were especially worried that insufficient time spent or messy drafts could negatively influence the teachers' perception on their writing.

Table 2.2: Topics identified in the bachelor students focus group (N = 5)

Topic	Sticky note example	Who	Number of sticky notes (participants)	Number of votes (participants)
Typing patterns	Information on typos, grammatical issues	S, T	7 (3)	3 (3)
Planning	What my first ideas were, before I started, to understand it better	T	6 (3)	3 (3)
General structure	Number of words	S	3 (2)	3 (3)
Draft version	The parts (paragraphs) that I struggled with	NT	3 (2)	2 (2)
Sentence structure	How often do you change your sentences	S	3 (2)	2 (2)
Analysis of text	Which mistakes are being made the most	T	8 (2)	1 (1)
Register	Highlight words that are part of the academic register	S	2 (1)	1 (1)
Time	How much time I spend making on making my assignment	NT	4 (3)	0 (0)
Errors and judgement	How many mistakes you make in one writing assignment	NT	3 (3)	0 (0)
No category	Summary graphs, statistics	S	1 (1)	0 (0)

Note. 'Who' indicates who should have access to the data: T = teacher, S = student, NT = not teacher.

PhD students. The PhD students wrote 36 ideas and categorized those into 8 topics (two ideas were uncategorized; Table 2.3). Time or productivity were the central themes addressed at several points throughout the discussion. Students were interested in how they could reduce "*staring at an empty screen time*" and whether it would be possible to predict the best time of the day to write. The main goal of this was to be more productive or postpone less. The PhD students considered that information on time and productivity should not only be available to themselves, for time self-regulation, but also to their supervisors. This was stated by one PhD student as follows: "*I would really like my supervisors to be able to help me to produce something earlier*". However, some students disagreed with that viewpoint. They preferred to not disclose to their supervisors how much time they spent, or whether they wrote in the middle of the night, because they did not want to get criticized on this "*unhealthy work habit*".

In addition, the PhD students were interested in their revision behavior, by detailing where and when they revised. In particular, they were interested in how feedback and comments from supervisors or reviewers affected their writing, both positively and negatively. Some PhD students argued they did not want to disclose this information to their teachers. A student stated "*I do not want [my supervisors] to know that I don't agree with what I'm writing [...] sometimes you just want to please your supervisor. – I specially have this with reviewers as well*". Again, some PhD students disagreed and said they had nothing to hide.

Table 2.3: Topics identified in the PhD students focus group (N = 5)

Topic	Sticky note example	Who	Number of sticky notes (participants)	Number of votes (participants)
Planning time	What time of day do I do my best writing and is this consistent/predictable for me	S, T	2 (1)	4 (4)
Number of revisions	Number of revisions per paragraph	NT, S, T	8 (4)	2 (2)
Supervisor effect	How affected am I by previous reviews	NT, T	6 (3)	2 (2)
Empty screen time	Total time spent in general (staring at it)	NT	5 (4)	2 (2)
Speed of writing	How much time did I spend on each sentence	S, T	4 (4)	2 (2)
Revision content	Type of revisions in terms of sentence structure or paragraph structure, content	S	3 (3)	2 (2)
No category	How did I decide what info made it to the text (esp. literature section)?	T	1 (1)	2 (2)
Planning structure	In which order do I write the typical research paper?	S, T	5 (5)	1 (1)
No category	Style/discipline style. More or less passive.	T	1 (1)	1 (1)
Change over time	Change in writing over time, does it speed up or improve	S	1 (1)	0 (0)

Note. 'Who' indicates who should have access to the data: T = teacher, S = student, NT = not teacher.

Teachers. The teachers wrote 37 ideas and categorized those into 10 topics (see Table 2.4). They provided detailed headers and, accordingly, most topics were discussed only once. The teachers were mostly interested to show students information on their language, especially regarding style or "*language that is not necessarily incorrect*". For example, feedback could be provided on how to improve the text by making the language more formal or using a wider variety of sentence structures. They stressed that this feedback should not be directive, but rather should focus on what could be improved. In this way, students still need to think about how to improve the language and style.

The teachers were interested in improving their own instruction regarding the extent of linearity of writing. For example, teachers would like to know in what order the different sections were written by the students. Additionally, they wanted to gain understanding about how feedback, and specifically peer feedback, can play a role during revision. One teacher suggested that it would be useful to reflect on evidence to answer the following: "*How do students use feedback to revise their work? Do they go through comments one by one, or do they focus on one type of error comment?*". In addition, the depth of students' revisions was highlighted, such as the significance of the changes and the difference between language versus content revisions.

Table 2.4: Topics identified in the teachers focus group (N = 4)

Topic	Sticky note example	Who	Number of sticky notes (participants)	Number of votes (participants)
Making changes in language	Show instances where the student can improve language that is not necessarily incorrect	S	7 (3)	3 (3)
Linearity of the writing process	How linear their writing process is	T	3 (3)	3 (3)
Use of feedback during the revision stage	To what extent students make significant changes to draft versions	T	5 (4)	2 (2)
Just do it!	Never start with the opening sentence. Just write!	S	5 (2)	1 (1)
Making changes in content	Where and how their argument works/does not work	S	3 (3)	1 (1)
Students' perceptions of writing	Which aspects of writing they enjoy	T	3 (2)	1 (1)
Using peer feedback	How to give correct feedback on texts of other students	S	1 (1)	1 (1)
How students start the writing process	Do writing plans really help the student?	T	6 (4)	0 (0)
The pre-writing stage	How to start writing. E.g., finding resources and how to start	S	3 (2)	0 (0)
Using genre conventions	How would students define an academic writing style?	T	1 (1)	0 (0)

Note. 'Who' indicates who should have access to the data: T = teacher, S = student.

Professional development staff. The professional development staff wrote 46 ideas spread over 11 topics (Table 2.5). A main theme in the first two topics was source-based writing, or how students use information in their writing (using evidence). These topics were also highly related to reading. For example, a staff member raised the following question that would ideally be desirable to be addressed with evidence: "*What kind of information do students extract from literature and how do they extract this?*" The professional development staff would like to provide this information to students, to show them how to map their evidence, and how to use resources judiciously; but also, to teachers, to determine whether students needed additional instruction. For example, a staff member suggested this could be achieved by providing them workshops on reading into writing to "*scaffold the reading, evaluating, and synthesizing processes*". The concepts of reading into writing and using evidence were also related to plagiarism. A staff member mentioned: "*We assume that everyone is going to draw on published readings for assessments in some way, or readings provided by the lecturer, but I want to know, what else are they using?*". In addition, the professional development staff highlighted the critical role of metacognition for students to understand the processes involved when writing. They would like students to know whether they are on the "*right track*" or provide information on what steps they need to go through when writing an assignment.

Table 2.5: Topics identified in the professional development staff focus group (N = 6)

Topic	Sticky note example	Who	Number of sticky notes (participants)	Number of votes (participants)
Reading into writing	How do students integrate the literature/evidence?	T	7 (6)	7 (6)
Using evidence	How to synthesize evidence sources into their writing?	S	4 (4)	3 (3)
Understanding the task	Is the student answering the question in the assignment?	S	2 (1)	3 (3)
Understanding the process	What steps do I go through when writing an assignment?	S	8 (5)	2 (2)
Planning, organizing ideas	How students' ideas/concepts develop as they plan, draft and revise?	T	2 (2)	2 (2)
Time	How long does it take to write an assessment from start to finish?	T	4 (2)	1 (1)
(Pre-)reading	What do students read and how do students read?	T	5 (4)	0 (0)
Editing	Where they need to proofread and edit?	S	4 (3)	0 (0)
Writing, drafting, revising process	When does most in depth revision occur? (between first draft, second draft, etc.)	T	4 (3)	0 (0)
Structure	How they approach the structure of the writing, i.e. which section do they address first?	T	3 (3)	0 (0)
Appropriate use of resources	What kinds of readings/texts are students using (e.g. published or other students' assessments)?	T	1 (1)	0 (0)
Planning	How they plan or didn't, e.g., overall structure of the text	S	1 (1)	0 (0)
Addressing the task	The extent to which students understand and address the assignment questions	T	1 (1)	0 (0)

Note. 'Who' indicates who should have access to the data: T = teacher, S = student.

Writing researchers/specialists. The writing researchers generated 22 ideas, grouped into 7 topics (one idea was uncategorized; Table 2.6). First, they would like teachers to know where students struggle during the writing assignment. This idea was rather clear for all researchers and only discussed briefly.

Second, time was a recurring theme during the discussions. Time was discussed in terms of duration, or the time spent on the assignment, but also in terms of the order of the different activities during writing, such as when students think and reflect on their writing. In addition, the periodicity of writing was discussed. "*Did they write everything at once, or in regular or irregular chunks spread over a period of time?*".

Third, researchers were interested to show information to students on their revisions; whether these are good enough to improve writing quality. The main goal was to encourage students to engage in critical thinking, to "*help students write more critically rather than descriptively*" or to simply think or revise more, or to revise at deeper levels. This was also related again to the time spent on writing. A researcher stated the following: "*Give [the students] a little bit of information on how much time they spent and how much time the other students are spending. And then suggest them to reflect on what they have written so far*".

2.3.2 Comparing topics and ideas across stakeholders

To compare and contrast the topics across stakeholders, the topics were mapped into the *planning*, *translating*, *reviewing* and *monitoring* processes. In addition, they were ordered in terms of the level at which the indicators were described: *low-level behavioral*; *behavioral in relation to time (ordering, scheduling) or the writing process*; or *higher-level cognitive*. Figure 2.1 depicts the mapping of the topics and shows several similarities and differences between the topics discussed. All stakeholders discussed the three writing processes: planning, translating, and reviewing as well as the monitoring process. In addition, all stakeholders emphasized time. Lastly, all focus groups discussed topics related to the task environment and discussed both behavioral and cognitive indicators of the writing process.

Table 2.6: Topics identified in the writing researchers/specialists focus group (N = 4)

Topic	Sticky note example	Who	Number of sticky notes (participants)	Number of votes (participants)
Being stuck	Where students struggle, e.g. answering the question, formulating an argument	T	3 (3)	3 (3)
Time	How much time students spend on a particular assignment	T	3 (3)	3 (3)
Revision types, levels	Suggested revision types and examples in their own text of where they've done them	S	3 (2)	3 (3)
Planning	Did they plan well enough before starting to write?	S, T	3 (2)	1 (1)
Macro edits	Moving text to build cohesion	S, T	4 (2)	1 (1)
No category	Their writing behavior/patterns + how this can be improved	S	1 (1)	1 (1)
Critical thinking	Where the core thesis changed substantially	S	4 (2)	0 (0)
Academic integrity	Alerts to large chunks of text posted in from somewhere - cheating?	T	1 (1)	0 (0)

Note. 'Who' indicates who should have access to the data: T = teacher, S = student.

Several differences were found across groups:

- First, some stakeholder groups focused more on behavioral indicators (first row in Figure 2.1), while others focused more on cognitive indicators. Bachelor students discussed mostly low-level behavioral indicators (e.g., number of keystrokes). PhD students also discussed behavioral indicators, but usually in relation to scheduling time or the writing process (e.g., what is the best time of the day to write). Teachers, writing researchers, and, especially, professional development staff discussed higher-level cognitive indicators, such as the understanding of the writing process or critical thinking.
- Second, the different aspects of the task environment (e.g., task sources or collaborators/critics; rectangular boxes in Figure 2.1) were not discussed by all groups. For example, the task description was only discussed by the professional development staff, while the text produced so far and collaborators/critics were only discussed by bachelor students, PhD students, and teachers.
- Lastly, some stakeholders identified that a certain topic would be only of interest for either students or teachers (indicated by an S or T in Figure 2.1, respectively), while others did not make a clear distinction: the topics were considered of interest for both. For example, the professional development staff thought it would be useful for students to know whether they *understood* the task, while it would be of interest for teachers to know whether the students *addressed* the task.

A closer look into the discussions revealed one additional key difference between the stakeholders in terms of the terminology each group used. These differences were especially found in discussions around time, planning, and revision. For example, all stakeholders discussed time in terms of *duration*, or how long it took to write. However, while most stakeholders reported duration was something the teachers should see, bachelor students specifically stated that teachers should not see this. All stakeholders except from teachers discussed time in terms of the *time until the deadline* or when the student started to write. All stakeholders, except from bachelor students, discussed time at a deeper level. On the one hand, teachers and writing researchers mentioned the *ordering of the writing process*, such as at what points in time students stop to reflect and revise, and the *ordering of the writing product*, such as which paragraph was written first. On the other hand, PhD students focused on the *scheduling writing* during the day and over multiple sessions.

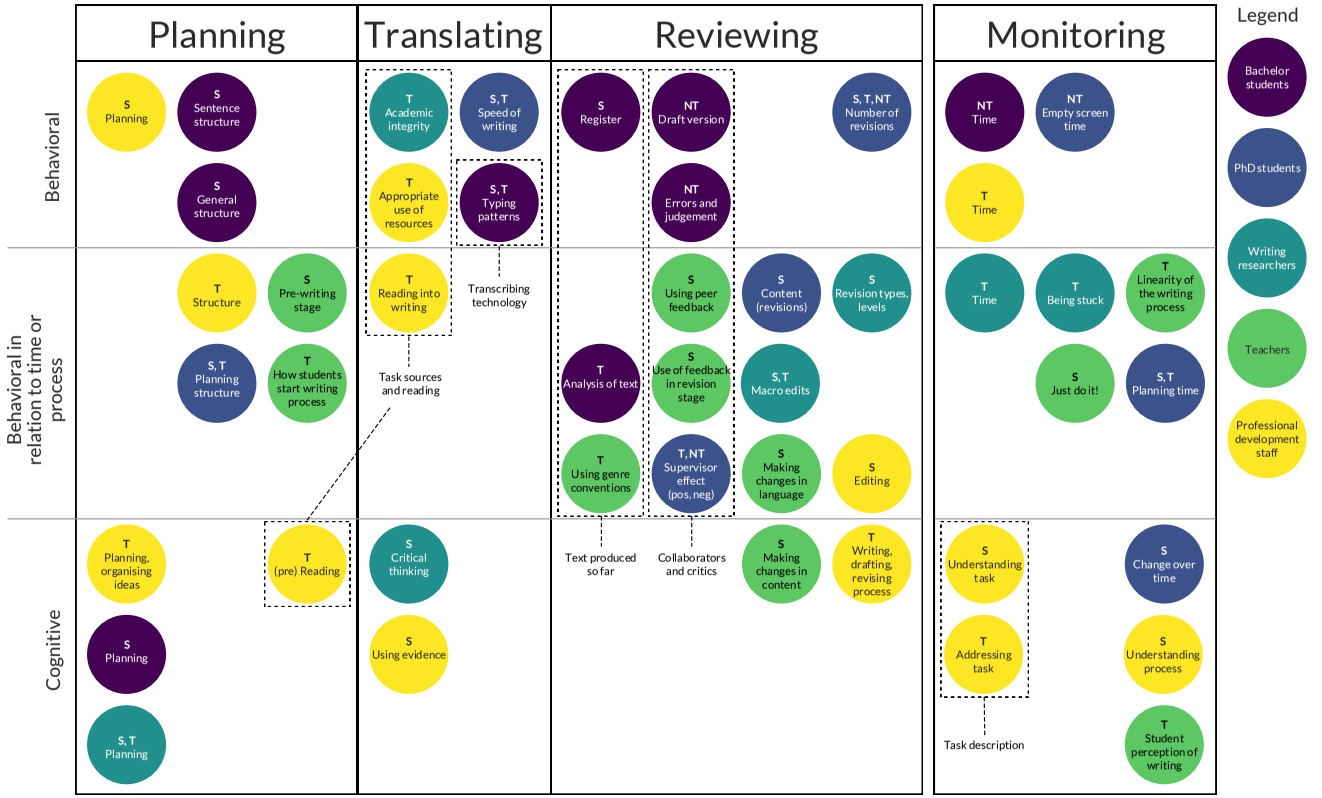

Figure 2.1: Overview of the topics discussed by all stakeholders.

Note. Topics are mapped to the process they address in Flower and Hayes' (1981) model of writing processes (x-axis) and level of the indicators (behavioral versus cognitive; y-axis). Topics related to concepts in the writing environment (Hayes, 2012) are grouped in the rectangular boxes. Letters in the circles (T = teacher, S = student, NT = not teacher) indicate the stakeholders who should have access to these data.

Likewise, different conceptualizations and properties of planning and revision were discussed. Planning was discussed in terms of planning structure, content, or language use, where *planning structure* was most often discussed. *Planning content* was only discussed by PhD students, teachers, and professional development staff, while *planning language* was only mentioned by professional development staff and bachelor students. Revision was discussed in terms of the different characteristics of revision. *Depth of revision*, such as surface-level versus structure or document (deep-level) changes was heavily discussed by all stakeholder groups. Other properties of revision included: the *temporal location of revision*, when revisions were made (PhD students, professional development staff, writing researchers); the *spatial location of revision*, such as which parts have been revised (PhD students, writing researchers); the *quality of revision* (professional development staff, writing researchers); and the *order of revisions* (teachers).

2.3.3 Integration into the learning design

After identifying the desired indicators for the stakeholders, we examined how these indicators could be integrated into learning and teaching practices, by designing use cases. Interestingly, most stakeholders within each focus group choose the same or a similar idea to integrate into the learning design. The use cases showed that the tools should not 'fix' the problem, but rather advise or suggest strategies to address it.

Specifically, professional development staff would like a tool to help students during reading, for integrating resources in their writing, and for synthesizing evidence. This tool would need to automatically pop-up during reading and writing, and help students by scaffolding reading into writing, with models, examples, guidelines, and strategies. It needed to be tailored to the disciplinary context, and students might actively choose what they want help with, and what kind of text (discipline) they are reading. The writing researchers proposed a similar tool, to help students critically reflect on their text. A message would pop-up when few or only low-level revisions are made or after a long time of inactivity. The tool would address what could be improved by using examples (from their own writing) and encourage students to critically reflect on what they wrote.

The tools envisioned by the bachelor students and teachers were focused on lower-level aspects of writing. Bachelor students envisioned a tool similar to a spell-checker, built in into their word processing software. The tool would flag incorrect referencing formatting, suggest words if the student is struggling finishing a sentence, and suggest synonyms if a

word is not from the academic register. Teachers came up with a similar tool, to flag informal words and suggest more formal words. This way, students would spend less time on these lower-level aspects of the text and would have more time left for structuring their argument. PhD students would like to have a dashboard, which keeps track of their productivity and number of revisions per section, for each writing session. This dashboard would be used before a new writing session, to identify the most productive time of the day, which section needs more attention, or the best time when to take a break.

Regarding the tools for teachers, professional developments staff would like to show videos of an expert's writing process to first-year students, to show how ideas develop over time. This would be used for workshops and instructions (face-to-face or blended) on strategies for approaching and scaffolding reading and writing. Another tool mentioned would measure the amount of critical reflection. This would be used to inform instruction, by explaining how to critically reflect within the specific discipline, by using models and examples.

2.4 Discussion

In this chapter we aimed to determine the indicators of higher education students' writing processes that are desirable to provide automated, personalized writing feedback for, and how this could be implemented into the learning design. These indicators were elicited and use cases for these indicators were developed through participatory sessions with bachelor students, PhD students, teachers, professional development staff, and writing researchers. All groups noted a variety of indicators which were grouped into self-generated categories. We mapped these categories into the planning, translating, reviewing, and monitoring processes as described by Flower & Hayes (1981). In addition, we coded the level of the indicators, ranging from low-level behavioral to higher-level cognitive indicators. This classification proved to be useful to compare and contrast the ideas between the different stakeholders and resulted in four main findings with implications for both writing tool design as well as writing process research.

First, we identified which indicators are desired by different stakeholders for providing automated feedback on the writing process. All stakeholder groups identified features in each of the major writing processes: planning, translating, reviewing, and monitoring, as described by Flower & Hayes (1981). Desired indicators for each of these processes respec-

tively, were, for example, information on students' planning strategies, how students used evidence in their writing, the depth of revisions, and students' understanding of the task.

Second, we showed that the level at which the indicators were discussed varied between the five stakeholder groups. These findings corroborate previous literature, which also indicated that students and teachers differ in their perceptions of academic writing (Itua et al., 2014; Lea & Street, 1998; Wolsey et al., 2012). Students focus more on lower-level indicators such as content and knowledge (Norton, 1990), while teachers focus more on higher-level indicators, such as argument and structure (Lea & Street, 1998; Norton, 1990). However, these previous studies mostly determined differences in perceptions of writing in relation to the writing product. In the current chapter, we showed that these differences also hold for perceptions of the writing process. Bachelor students focused on rather low-level behavioral indicators of the writing process, such as the number of keystrokes. By contrast, teachers, writing researchers, and especially the professional development staff focused on higher-level cognitive indicators, including critical thinking and the understanding of the writing process.

Third, extending on previous work which identified two levels at which indicators were discussed (Lea & Street, 1998; Norton, 1990), we distinguished a third (intermediate) category, in which behavioral indicators were discussed in relation to time or the writing process. Researchers have argued that time needs to be considered when studying writing processes, as it might provide information regarding the purpose of a specific processes and how sequences of cognitive processes differ across writers (Rijlaarsdam & Van den Bergh, 1996). For example, both novice and expert writers might show the same frequency of cognitive activities, but expert writers might know *when* they need to engage in which activity. Indeed, we found that PhD students more often discussed behavioral indicators in relation to time, e.g., what is the best time of the day to write, compared to bachelor students. This indicates that bachelor students, to become more expert writers, might need more active instruction to consider their writing actions in relation to time and the writing process.

Fourth, the results from the use cases provided some initial ideas on what tool sets, strategies, and actions are needed by whom and when, to integrate feedback on the elicited indicators of the writing process into the learning design (see also Section 2.4.1). Overall, these findings validate the usefulness of engaging multiple stakeholders in the identification of the key metrics that could be included into the design of writing analytics tools.

2.4.1 Implications for writing tool development

Currently, many writing tools solely provide summative and formative feedback on the writing product, rather than the writing process (Allen et al., 2015). Our findings provide insight into desirable features to extend these tools with indicators of the writing process. For example, information on students' planning strategies or the depth of revisions could be used to support students' reflection or for teachers to provide more effective feedback. To achieve this, tools could suggest strategies to encourage students to address the problem and develop their own writing strategies.

In addition, our findings provide implications for the design of the writing tools. We found differences in the terminology used by different stakeholders. Moreover, differences were found in what indicators would be useful for students and what indicators would be useful for teachers. These differences indicate that a user-centered approach needs to be taken to develop writing tools, in which either a common language need to be created to talk about writing processes, or in which different interfaces are created for different stakeholders (Gabriska & Ölveckỳ, 2018; Teasley, 2017). In addition, this indicates that students might need additional explanations to understand the higher-level aspects of the writing process. These explanations can come from the teachers (e.g., face-to-face or blended, in combination with the writing tool) or might be automatically triggered. Previous work already showed that feedback related to specific parts in the student text ('specific feedback') is more effective and requires less mental effort compared to general feedback (Ranalli, 2018). Hence, to provide better explanations of the writing process, it might be good to tie the feedback to specific examples in the writing product. All these differences in perspectives of the stakeholders further highlight the need for a human-centered approach (Giacomin, 2014) and hence the need for stakeholder involvement in the development of writing tools.

Finally, the use cases provide some practical implications for the (further) development of writing tools and the integration into the learning design. First, the tool should be tailored to the disciplinary context. The tool should (preferably automatically) detect a problem, but should not fix the problem, rather, it should provide instruction to address it. Professional development staff still preferred this instruction to be face-to-face or, if necessary, blended, indicating the importance of instructors' willingness to adopt the technology into the classroom (cf. Link et al., 2014). In addition, this indicates that the resulting

intervention should include both the detection of the problem (with the tool), as well as instruction (with or without the tool). This further stresses the claim made by Wise & Jung (2019) who also indicated the importance of studying how tools are used in real educational contexts.

2.4.2 Implications for writing process research

The indicators identified in this chapter have important implications for writing process research. Several of the indicators identified by the stakeholders have already been extracted by keystroke analysis. This specifically holds for the lower-level behavioral features, such as the number of keystrokes (e.g., Allen, Jacovina, et al., 2016), total time spent writing (e.g., Bixler & D'Mello, 2013), and the number of characters that stayed in the final product (e.g., Van Waes et al., 2014). However, this study showed that for providing automated and personalized feedback, it is critical to extract these behavioral indicators in relation to time or when they happen in the writing process, such as the order in which errors are revised or how the writing fluency changes over time. To date, little work has examined the temporal aspects of the keystroke data, with some exceptions (Likens et al., 2017; M. Zhang et al., 2016). Therefore, we suggest future work should focus on sequence mining and temporal analysis of the keystroke data, rather than solely extracting frequency metrics.

We also showed that higher-level cognitive features are considered desirable for providing feedback, such as how students synthesize evidence sources into their writing or how their ideas and concepts develop over time. Some indicators might not be accessible via keystroke data, such as the ideas students had before writing. For such indicators, think-aloud or structured reflection and planning tasks might be more suitable methods. To further fill the gap between keystroke data and cognitive processes, and especially to provide feedback, future work should investigate these data in combination with other sources of contextual information (Galbraith & Baaijen, 2019). For example, natural language processing on the text composed during the writing process in combination with temporal analysis could be used to extract different features related to revision, which could indicate the depth, timing, and location of the revision (see e.g., F. Zhang & Litman, 2015).

2.4.3 Limitations

The findings in this chapter are limited in two ways. First, we only analyzed five stakeholder groups. Within these groups, all students came from similar disciplines, while the

teachers had different backgrounds. Disciplinary background has shown to have an impact on teachers' opinions on most important elements of students' writing (Lea & Street, 1998) and on students' conceptions of essay writing (Hounsell, 1984, 1997). Therefore, additional focus groups with different disciplines could have resulted in more and other indicators. However, we did not aim to provide a full overview of all indicators desirable for providing feedback. We rather showed how a participatory approach could provide insight into what types of indicators are considered useful and how this could be integrated into the learning design. A possible future step in the design process would be to feed these insights back to the stakeholders, to comment on each other's insights and close the feedback loop.

Second, we focused on indicators that would be considered *desirable* to provide automatic and personalized feedback. However, desired indicators are not necessarily technically feasible or useful indicators. Future work needs to determine which indicators can actually be extracted (see also Section 2.4.2). In addition, the indicators do not necessarily improve writing proficiency and might not even have an impact on writing quality of a specific writing product. Although several studies have shown that indicators of the writing process have a relation with writing quality (e.g., Allen, Jacovina, et al., 2016; Xu, 2018) and several writing tools have shown to improve motivation or writing quality (Cotos, 2015), the evidence is still limited and usually generalized over a whole tool, rather than for specific indicators. Therefore, future (empirical) studies are necessary to determine whether these indicators can positively impact writing and how these should be integrated into the learning design to positively impact writing.

2.5 Conclusion

In this chapter, in contrast to post-hoc user-centric evaluations of specific writing tools conducted after the development, we presented a participatory approach that happened before the development. Through an illustrative study, we showed which indicators are considered desirable by students, teachers, writing researchers, and professional development staff to provide automated, personalized writing feedback in higher education. Bachelor students focused mostly on lower-level behavioral indicators and PhD students mostly on behavioral indicators in relation to time, while teachers, writing researchers and professional development staff focused more on higher-level cognitive indicators. These lower-level behavioral indicators can be extracted automatically using keystroke logging. To go

beyond these lower-level features, writing process research should focus on temporal analysis of keystroke data and natural language processing of the text written so far, to gain a better understanding of the relation between keystroke data and cognitive writing processes. In addition, future work should further analyze how information on the writing process may be incorporated into writing tools and the learning design. We showed how stakeholder involvement in the form of a participatory approach can be valuable to further this goal.

3

The effect of writing task on keystroke data

Adapted from: Conijn, R., Roeser, J., & van Zaanen, M. (2019). Understanding the keystroke log: The effect of writing task on keystroke features. *Reading and Writing*, *32*(9), 2353–2374. https://doi.org/10.1007/s11145-019-09953-8

Keystroke logging is used to automatically record writers' unfolding typing process. However, it is not clear which and how features from the keystroke log map to higher-level cognitive writing processes, such as planning and revision. In this chapter we aim to investigate the sensitivity of frequently used keystroke features across tasks with different cognitive demands. Two keystroke datasets were analyzed: one consisting of a copy task and an email writing task, and one with a larger difference in cognitive demand: a copy task and an academic summary task. The differences across tasks were modeled using Bayesian linear mixed effects models. Posterior distributions were used to compare the strength and direction of the task effects across features and datasets. The results showed that the mean of all interkeystroke intervals were found to be stable across tasks. Features related to the time between words and (sub)sentences only differed between the copy and the academic task. Lastly, keystroke features related to the number of words, revisions, and total time, differed across tasks in both datasets. To conclude, our results indicate that the latter features are related to cognitive load or task complexity, and hence might be used to gain insight into students' writing processes. In addition, the findings show that keystroke features are sensitive to small differences in the writing tasks at hand.

Acknowledgements. I would like to thank Diana Schmalzried and Xinran Wang for their assistance in collecting the academic writing dataset.

3.1 Introduction

Insight into students' writing processes can provide evidence on where and when students struggle (Likens et al., 2017) and could be used to improve their writing ability (Deane, 2013). However, writing is a complex, non-linear process, where different cognitive processes interact and can happen in any order. Flower & Hayes' (1980) model distinguishes three main cognitive processes that interact: planning, translating, and reviewing. Given this complexity, it is difficult to provide automated methods that allow insight into students' writing processes (Baaijen et al., 2012; Leijten & Van Waes, 2013).

In the current thesis, we focus on the use of keystroke logging to measure writing processes during typing. The analysis of these keystroke logs, keystroke analysis, is a promising area of research, because keystroke logs provide real-time, fine-grained information on writers' unfolding typing process during text composition. In addition, keystroke logs can be collected automatically, hence keystroke logging is more scalable and less intrusive than traditional thinking-aloud methods and observation studies. Keystrokes have been used for a wide range of studies (e.g., Grabowski, 2008; Van Waes et al., 2014; M. Zhang et al., 2016), which used a variety of keystroke features to answer their research questions. However, it is not yet clear how each of these features map onto underlying cognitive processes, which has been coined as the problem of alignment (Galbraith & Baaijen, 2019). Keystroke features may be multiply determined; they are sensitive to a variety of factors. In addition, keystroke features are not independent, and will, at least to some extent, overlap in the cognitive processes they are representing. Therefore, it is not always clear which features need to be selected for the question at hand. The selection of the correct keystroke features is crucial for the interpretation of the results and therefore, the derived conclusions. Hence, there is a need for a better understanding of sensitivity and independence of the keystroke features frequently used in the writing literature. In this chapter we investigate the sensitivity of keystroke features across tasks.

3.1.1 Features extracted from keystroke logs

The features extracted from keystroke logs in previous work can be broadly organized into three categories: (1) features related to the duration of the keystrokes, (2) features related to content or revising behavior, and (3) features related to written language bursts (e.g., Baaijen et al., 2012; Bixler & D'Mello, 2013). In the majority of studies the researchers

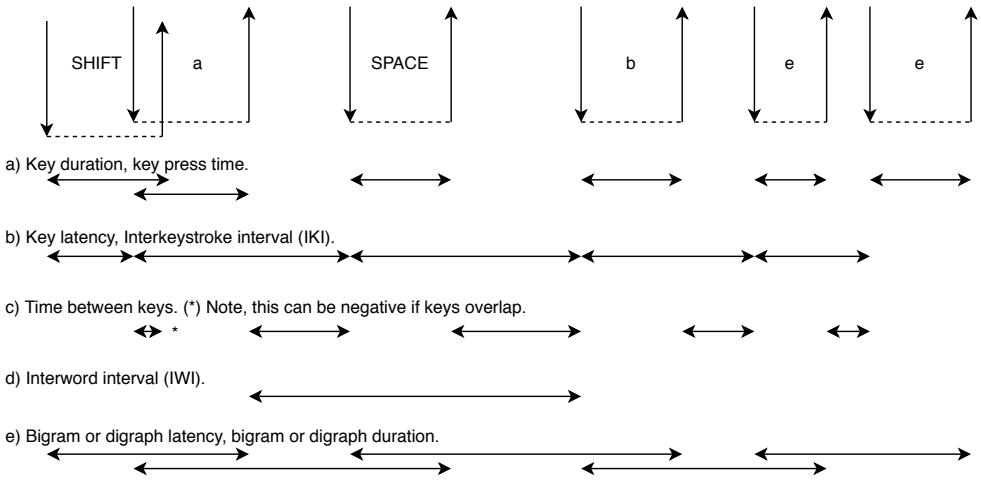

Figure 3.1: Time-based features extracted from the keystroke log of typing: "A bee".

extracted at least one or a few features related to duration, such as the duration between two consecutive key presses (e.g., Salmeron-Majadas et al., 2014) or the duration of one key press (e.g., Allen, Mills, et al., 2016; Bixler & D'Mello, 2013). The terminology of these time-based features is sometimes used interchangeably. For clarity, we provide an overview of the time-based features which are extracted from keystrokes (see Figure 3.1).

The specific duration features used depend on the hypotheses of the studies. For example, a literature review by Karnan et al. (2011) showed that the majority of writer identification and authentication studies focus on features such as key duration, keystroke latency, and digraph latency (see Figure 3.1a, b, e). Sometimes the time between keys (see Figure 3.1c) was included as well (e.g., Tappert et al., 2010). For these features several summary statistics were computed, such as the mean and the standard deviation of the keystroke latencies. Also vectors representing keystroke features have been compared using Euclidean distance (Giot et al., 2009). All these measurements were computed for specific keys (e.g., "b"), and for specific combinations of keys (e.g., the bigram "be"). These combinations indicate how much time it takes to type a specific key, or sequence of keys.

By contrast, writing analytics studies focus on timing features in general, and not related to a specific key. The interkeystroke interval (IKI, see Figure 3.1b) is the most commonly used feature. Multiple summary statistics are derived from IKIs, including the

largest, smallest, mean, and median IKI. Note that these summary statistics are not merely used to describe the data, but as outcome variables representing the cognitive process of interest. Sometimes, these statistics are calculated per pause location; for example the IKI within or outside words (Grabowski, 2008), or the IKI within words, between words, between subsentences, and between sentences (Baaijen et al., 2012). Alternatively, frequencies of IKI of specific lengths are extracted, such as the number of IKIs between 0.5–1.0 seconds, 1.0–1.5 seconds, 1.5–2.0 seconds, 2.0–3.0 seconds, and $>$ 3.0 seconds (Allen, Jacovina, et al., 2016). In addition, other features related to time are extracted, such as initial pause time (e.g., Allen, Jacovina, et al., 2016), interword interval (e.g., M. Zhang et al., 2016), function or content word time (Banerjee et al., 2014), and total time (e.g., Bixler & D'Mello, 2013).

Aside from duration features, content-related features are extracted, such as the number of keystrokes or verbosity (e.g., Allen, Mills, et al., 2016), the number of alphabetical keystrokes (Salmeron-Majadas et al., 2014), the number of words (e.g., Likens et al., 2017), the number of backspaces/deletes (e.g., Allen, Jacovina, et al., 2016), and efficiency, the number of characters in the final product per number of characters typed (Van Waes et al., 2014).

Lastly, some studies included features related to written language bursts. Writers compose sentence parts, identified by latencies longer than two seconds or by a grammatical discontinuity (Kaufer et al., 1986), also known as bursts. In keystroke analysis, written language bursts are often operationalized as sequences of text production (keystrokes) without an IKI longer than two seconds and without a revision and without an insertion away from the leading edge (Baaijen & Galbraith, 2018). Features related to written language bursts include the number of bursts or the number of words per burst after a pause, revision, or insertion (Baaijen et al., 2012).

3.1.2 Rationales for keystroke feature selection

The scientific rationales for selecting keystroke features can be divided into data-driven and theory-driven approaches. Studies using keystroke analysis for authentication and identification can be considered data-driven. These studies use multiple duration and content features to understand to what extent or with which accuracy the writer can be predicted, for example to build accurate automatic detection systems. Since including more features could lead to higher accuracy, often a combination of features is used that are known for

their predictive power from previous studies, and 'new' features that are hypothesized to have predictive power (e.g., Bixler & D'Mello, 2013; Karnan et al., 2011). Since the main focus is on predictive accuracy, understanding the relation between the keystroke features and the cognitive processes is of limited interest in these studies. However, a large information gain of a keystroke feature on the prediction, found in multiple studies or contexts, may indicate a relationship worth investigating.

On the contrary, there are theory-driven approaches for keystroke feature selection. Several studies link the keystroke features to the three writing processes as defined by Flower & Hayes (1980): planning, translating, and reviewing processes. In addition, keystroke features have been related to cognitive load. Cognitive load reflects the notion that task performance is bound by the working memory capacity available for cognitive processing and the cognitive demands of a task (Sweller, 1988). If the cognitive demands of a task exceed the available working memory capacity, the writer might slow down, other (less demanding) strategies might be used, or more errors might be made (Just & Carpenter, 1992). In writing, high-level processes such planning and reviewing, are considered to have a high cognitive demand as they require high levels of attentional control (Alamargot et al., 2007; Kellogg, 1996). By contrast, motor processes, such as typing, require less attention and hence have a lower demand (Olive & Kellogg, 2002).

Cognitive load, or planning and revising processes in general, are commonly related to duration features, such as the number of, length, and location of IKI or pauses (Van Waes et al., 2014; Wengelin, 2006). More pauses and longer pauses are related to a larger cognitive load (Alves et al., 2007; Wallot & Grabowski, 2013), that, for example, could indicate word and sentence planning or deliberation (Roeser et al., 2019; M. Zhang et al., 2016). By contrast, shorter pauses are related to basic keyboard fluency or motor processes (Grabowski, 2008). Several studies also distinguish between different pause locations, such as between words or between (sub)sentences, or between and within words. Pauses before words are considered to reflect planning, retrieving, verifying, or editing processes, while pauses within words are considered to be related to typing skills (Baaijen et al., 2012; Grabowski, 2008). Pauses at sentence boundaries are considered to reflect global text planning and require more time, compared to pauses at word boundaries, which are considered to reflect lexical access (Medimorec & Risko, 2017). Features associated with content and revising content are frequently related to translation and revision processes. The number

of words is often related to writing quality, where more words indicate a higher essay quality (e.g., Allen, Jacovina, et al., 2016). The number of deletions is argued to be related to revision processes (Van Waes et al., 2014), but also to lower-level aspects such as keyboard efficiency (Grabowski, 2008). Lastly, written language bursts are related to the execution process or Flower & Hayes' (1980) translation processes (Baaijen et al., 2012). Longer bursts and shorter pauses have been related to higher writing fluency and improved text quality (Alves & Limpo, 2015).

Thus, keystroke features are used to infer variations related to cognitive writing processes and cognitive demand required for these writing processes. However, it is unclear how exactly the features are related to these cognitive writing processes and how sensitive the features are to differences in cognitive demand (Galbraith & Baaijen, 2019). As different writing tasks are bound to reflect different cognitive demands, investigating differences in keystroke features across tasks may provide insight into the sensitivity of keystroke features to differences in cognitive demands.

3.1.3 Sensitivity of keystroke features across tasks

Previous studies showed that keystroke features differ between tasks. For example, features were shown to differ between a copy task (transcribing a text) and a—more demanding—email writing task (e.g., Tappert et al., 2010). However, these differences were not made explicit nor evaluated. Other studies did explicitly state the differences. Conijn & van Zaanen (2017a) found differences in the number of keystrokes, number of corrections, mean and standard deviation of IKI within, before, and after word between an email writing and a copy task. Grabowski (2008) added a third task: copy from memory. Here, copying from text was considered more difficult than copying from memory, because the former task also included eye-hand coordination, needed for reading and reproducing the text. The most difficult task was email writing which involves planning and formulation in addition to motor-planning and execution. Results showed a larger efficiency (ratio between the number of characters in the final document and the number of keystrokes) for copy from memory, compared to copy from text and generation from memory. Typing speed, measured by IKI between and within words, was found most stable across tasks. In the current chapter, we extend on this work by analyzing the differences in keystrokes across multiple tasks, which are assumed to differ in the required cognitive load and therefore, affecting keystroke features related to the cognitive processes involved.

3.1.4 Current approach

We aim to investigate which, and how, keystroke features are affected by differences in cognitive load across writing tasks. This provides insight into the sensitivity of these features and which features are useful for analyzing cognitive writing processes. Finally, this could be used by teachers or instructional designers to evaluate differences in cognitive demands imposed by their chosen learning designs.

Two datasets were collected, both containing keystroke data from two tasks: (1) Villani dataset, consisting of a copy task, where participants were asked to transcribe a given printed text, and an email writing task; and (2) Academic writing dataset, consisting of a copy task and an academic summary task. The copy task and the email writing task differ in terms of planning and revising processes. In a copy task, there will be no planning on a linguistic level, but only planning on a motor level (eye-hand coordination). In addition, revising will only take place for typos, but not for linguistic reasons. The copy task and the academic task differ even more in terms of planning and revising processes, compared to a copy task and an email writing task. This is because academic writing involves additional complexity, such as critical thinking, integrating sources, and utilizing a repertoire of linguistic practices appropriate for the task (Lea & Street, 1998).

Bayesian linear mixed models were used to determine the effect of these tasks on the keystroke features. Several keystroke features related to keystroke duration and deletions were extracted, because these have been related to cognitive load in general. We hypothesize that specifically features related to the time between words, the time between sentences, and the amount of revision are sensitive to the tasks, as these are well-documented in the literature to be associated with cognitive writing processes (Van Waes et al., 2014; Wengelin, 2006). By contrast, we hypothesize that features related to keystroke duration within words are not sensitive to the tasks, because these have been associated with motor processes (Grabowski, 2008).

3.2 Method

Data were collected from two different datasets, both containing two different tasks: the Villani keystroke dataset, containing a copy task and an email writing task, and a dataset on academic writing recorded for the purpose of this research, containing a copy task and an academic summary task. The copy tasks differed to the extent that different texts were used

to transcribe. However, both copy tasks did not require higher-level cognitive processes, such as planning on a linguistic level, as involved in the other tasks.

3.2.1 Villani dataset

The Villani keystroke dataset (Monaco et al., 2012; Tappert et al., 2010) is an open-source keystroke dataset collected in an experimental setting. In the experiment, students and faculty could choose to conduct a copy task and/or an email writing task. The participants were allowed to type both forms of text multiple times. For the copy task, the participants were asked to copy a fable of 652 characters. In the email writing task, the participants were asked to write an arbitrary email of at least 650 characters. During the experiment, the key typed, time of key press, and time of key release were stored for every keystroke. In total, this resulted in more than one million keystrokes. The dataset consists of 142 participants, who wrote 359 copy texts and 1262 emails. The dataset and the collection of the dataset is explained in detail in Tappert et al. (2010).

For the current study, several data cleaning steps were taken. First, we only included data from participants who participated in both the copy task and the email writing task. This resulted in a dataset of 36 participants, who collectively wrote 338 copy texts and 416 emails. Second, inspections of the dataset showed some cases where a key was only released after a subsequent key was pressed, resulting in a negative time between keys. This for example happens when typing combination keys, such as SHIFT + {a–z} to capitalize a letter. Since we are interested in writing characteristics that differ across tasks, not in character-specific information such as capitalization, all times between keystrokes, words, subsentences, and sentences which were lower than 0, were coded as missing. Lastly, some participants typed only a few characters or clearly typed random sequences of characters, without spaces. Therefore, seven sessions were excluded where the number of keystrokes was smaller than 600 or the number of words was smaller than 50. This left us with a total of 747 sessions.

3.2.2 Academic writing dataset

The academic writing dataset was collected in an experimental setting; in an academic writing course for English second language learners. As part of the course, students were asked to complete two tasks: a copy task and an academic summary task. For the copy task, the students were asked to transcribe a fable of 850 characters. For the academic summary task,

the students were asked to write a summary of 100–200 words based on a journal article. The journal article (Woong Yun & Park, 2011) described a 2 × 2 experimental design in the field of the students' major (communication and information sciences). After reading the article, students were asked to write a summary within 30 minutes. During both tasks, keystrokes were collected using Inputlog (Leijten & Van Waes, 2013), from those students who provided informed consent. In total, 131 students participated in the study.

Similarly to the data cleaning of the Villani dataset, only data were included from participants who completed both the copy task and the summary task, resulting in data from 128 participants. In addition, all times between keystrokes, words, subsentences, and sentences which were lower than 0 were coded as missing. Lastly, since the summary task was considerably longer than the copy task, we only selected a subset of keystrokes of the summary task. Participants typed on average more than 900 characters in the copy task. Therefore, the first 900 characters were extracted from the summary task (session 1). If the participant wrote less than 900 characters, all characters were extracted. In addition, as most participants wrote more, the next 900 characters (901–1800) were also extracted from the summary task (session 2). For participants that wrote less than 1800 characters, all characters from character 901 were extracted, resulting in two subsets of keystrokes per participant. In addition, similarly to the Villani dataset, sessions were excluded where the number of keystrokes were smaller than 600 or the number of words were smaller than 50. This resulted in a total of 128 copy task sessions and 115 (session 1) + 67 (session 2) = 182 summary task sessions.

3.2.3 Feature extraction

From both datasets, we extracted frequency-based and time-based features similar to those used in writing analytics literature. Five frequency-based features were extracted from the task as a whole, related to content and revision behavior: number of keystrokes, number of words, number of backspace or delete keys pressed, efficiency (which is defined as the number of characters in the final document divided by the number of keystrokes), and the number of interkeystroke intervals (IKI) between 0.5 and 1.0 seconds. Although the number of IKIs larger than 1.0 seconds (IKIs between 1.0–1.5 seconds, 1.5–2.0 seconds, 2.0–3.0 seconds, and larger than 3.0 seconds) has been used as feature in previous writing studies (Allen, Mills, et al., 2016; Bixler & D'Mello, 2013), these were barely present in our dataset, and therefore not included in the present analysis.

Twenty time-based features were extracted. Seven of these were related to general keystroke durations, such as IKIs, the most commonly used feature in the literature, including mean, standard deviation, median, largest, and smallest IKI, as well as the mean and standard deviation of the key press time. Additionally, time-based features were extracted, which were related to specific locations in the text, including the mean and standard deviation of IKI between words, IKI within word, the time between keys, the time between words or the interword interval, the time between sentences (indicated by periods, question marks, and exclamation marks), and the time between subsentences (indicated by commas, semicolons, and colons, as in Baaijen et al., 2012). Lastly, the total time of the task was computed. The time-based features showed large variation. To account for this positive skew, all time features (except for total time) were log transformed and all values above the 95th percentile were removed. Similar approaches were used in previous studies (e.g., Grabowski, 2008; Van Waes et al., 2017). Figure 3.2 and Figure 3.3 show the distributions of each features extracted for each type of task in the two datasets, respectively.

3.2.4 Analysis of differences in keystrokes between tasks

Bayesian linear mixed effects models (BLMMs; Gelman et al., 2014; Kruschke, 2014; McElreath, 2016) were used to determine the differences in keystroke features between tasks within each dataset and across the datasets. All keystroke features were used as dependent variables, and task (copy versus email writing and copy versus academic summary for the respective datasets) was added as a fixed effect. Participant ID was added as a random intercepts term accounting for variance in the keystroke features specific to individuals. In addition, the effect of task on the keystroke features might differ across less-experienced and more-experienced writers, as hypothesized by Grabowski (2008). This possibility was accounted for by adding by-participant slopes for task.

In the context of this study, a Bayesian approach was chosen for three reasons. First, BLMMs provide a reliable way of accounting for differences related to participant and task, with guaranteed convergence (Bates et al., 2015). Second, BLMMs make it possible to derive posterior probability distributions of the variables of interest (here, the task effect for each keystroke feature). Lastly, these posterior probability distributions can be used to compare the effect across various dependent variables within a dataset and, more importantly, across datasets.

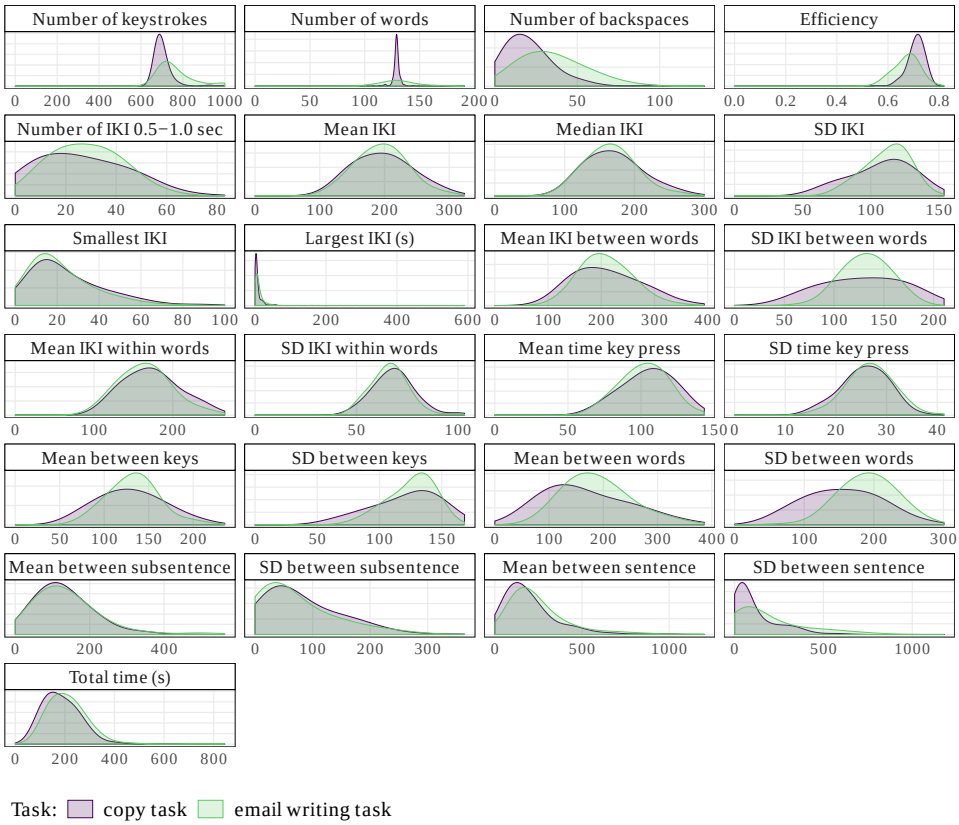

Figure 3.2: Distributions of the keystroke features per task (after trimming), for the Villani dataset.

Note. All times are in ms (except total time and largest IKI [in secs]). For visualization purposes only, values larger than 4000 ms for the mean and *SD* time between sentences in the academic writing dataset were removed.

For continuous models, linear models with log-normal distributions were used. For frequency data such as the number of words, distributions of the Poisson family were used. When discrete values were highly zero-inflated, e.g., included a large number of zero backspaces, negative binomials were used (Gelman & Hill, 2006; Gelman et al., 2014). Quasi-logit regressions were used for ratio data (e.g., efficiency), as these are bound between 0 and 1 (see e.g., Agresti, 2002; Barr, 2008; Donnelly & Verkuilen, 2017). In other words, the dependent variable was transformed from proportions to adjusted logits, and fitted as

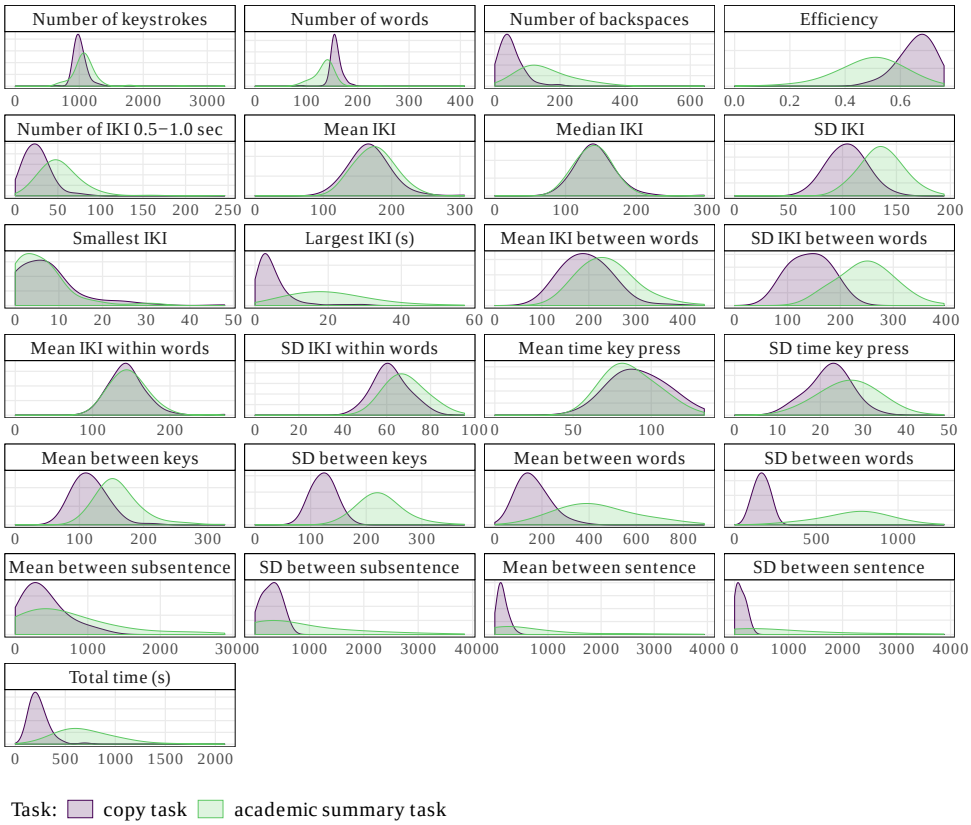

Figure 3.3: Distributions of the keystroke features per task (after trimming), for the academic writing dataset.

Note. All times are in ms (except total time and largest IKI [in secs]). For visualization purposes only, values larger than 4000 ms for the mean and *SD* time between sentences in the academic writing dataset were removed.

a continuous variable in linear regressions.[1]

The task effect (copy versus email/academic summary) on the keystroke features was evaluated in two ways. First, the most probable effect estimate $\hat{\beta}$ and its 95% credible in-

[1] BLMMs were conducted in R using the R-package "rstanarm" (Gabry & Goodrich, 2016). Weakly informative priors were used. The number of Markov chain Monte Carlo chains was set to 3 with 3,000 iterations per chain (1,500 warm-up). The Rubin-Gelman statistic (Gelman & Rubin, 1992), traceplots, and leave-one-out cross-validation were used to determine model convergence (Vehtari et al., 2015, 2017).

terval were calculated to determine the size and direction of the effect. In contrast to confidence intervals, credible intervals indicate the range in which the true (unknown) parameter value (here, the task effect) lies with 95% probability (Kruschke, 2014; Nicenboim & Vasishth, 2016; Sorensen et al., 2016). If a credible interval includes zero, zero is a possible estimate of the effect of task on the outcome variable (here, the keystroke features).

Second, the posterior probability distribution was used to calculate the standardized effect strength $\hat{\delta}$ which is defined as $\hat{\delta} = \frac{\hat{\beta}}{\hat{\sigma}}$, where $\hat{\beta}$ is the task effect estimate, and $\hat{\sigma}$ is the variance estimate for this effect. This effect strength allows us to compare the task effect across keystroke features within and across datasets, as it depends less on methodology or other experiment specific variables, such as language, text type, or participants, than the estimates (Wagenmakers et al., 2010).

3.3 Results

The results of the Bayesian linear mixed effects models show that task has an effect on several keystroke features (see Table 3.1). The direction of the task effect was largely similar across datasets. Specifically, most keystroke features showed a positive effect in both datasets, indicating larger values for the email writing task or the academic summary task, compared to the copy tasks. For example, participants paused 36 ms longer between words in the email writing task, compared to the copy task, and 208 ms longer between words in the academic summary task, compared to the copy task. Only efficiency and mean key press time were smaller for the email writing and the academic summary task. For the email writing task, efficiency was 41% lower and the mean key press time was 5 ms lower, compared to the copy task. For the academic writing dataset, efficiency was 19% lower and mean key press time was 13 ms lower, compared to the copy task. Thus, in the copy tasks, fewer keystrokes were needed per character in the final document and keys were pressed for a shorter period of time. In addition, the mean IKI within words and the smallest IKI were smaller for the email writing task, compared to the copy task, whereas the number of words and the mean IKI showed smaller values for the academic summary task.

Interestingly, for the *SD* of time between subsentences, and mean IKI within and between words the most probable effect value ($\hat{\beta}$) changed in direction. Specifically, the *SD* of time between subsentences and the mean IKI within and between words were lower for the copy task, compared to the email writing task in the Villani dataset. However, these were larger for the copy task, compared to the summary task in the academic writing dataset.

Table 3.1: Task effects estimated from BLMMs on keystroke features

	Villani dataset			Academic writing dataset		
Keystroke feature	Lower	$\hat{\beta}$	Upper	Lower	$\hat{\beta}$	Upper
Number of keystrokes	1.01	1.04	1.08	1.38	1.45	1.51
Number of words	-1.06	-0.99	-0.93	-0.78	-0.73	-0.68
Number of backspaces	1.02	1.12	1.22	1.27	1.35	1.42
Efficiency	-0.44	-0.41	-0.39	-0.22	-0.19	-0.16
Number of IKI 0.5-1.0 sec	1.11	1.51	1.97	4.69	8.46	12.88
Mean IKI	-11.77	-2.70	5.61	-16.54	-10.74	-5.21
Median IKI	1.02	1.12	1.22	1.27	1.35	1.42
SD IKI	0.06	0.11	0.16	0.52	0.60	0.69
Smallest IKI	-7.65	-4.26	-0.80	-2.45	-1.15	0.25
Largest IKI (s)	8.81	17.14	27.27	133.93	189.57	247.65
Mean IKI between words	-12.83	-0.19	12.43	20.31	27.13	34.08
SD IKI between words	0.05	0.11	0.18	0.63	0.71	0.80
Mean IKI within words	-18.54	-11.69	-4.52	-0.63	2.71	6.18
SD IKI within word	-0.00	0.04	0.09	0.02	0.06	0.09
Mean time key press	-7.54	-4.55	-1.49	-14.86	-12.64	-10.52
SD time key press	0.02	0.05	0.07	0.26	0.31	0.36
Mean between keys	1.87	10.12	19.04	19.72	25.38	30.66
SD between keys	0.01	0.11	0.20	1.26	1.48	1.70
Mean between words	19.16	36.31	53.54	172.40	208.28	246.29
SD between words	0.14	0.30	0.47	2.96	3.37	3.84
Mean between subsentence	-23.14	6.66	44.30	121.39	284.03	460.96
SD between subsentence	-0.39	-0.12	0.16	-0.32	0.44	1.19
Mean between sentence	16.35	74.10	138.60	470.47	1056.81	1778.19
SD between sentence	0.10	0.39	0.70	10.92	18.56	27.39
Total time (s)	51.56	80.05	108.14	2692.67	3295.20	3853.06

Note. All values are shown in their original units. All times are in ms (except from total time and largest IKI). Positive values indicate larger values for the email writing or academic summary task and negative values indicate larger values for the copy tasks. $\hat{\beta}$ is the most probable estimate for the difference between tasks. Lower and upper specifies the 95% credible interval around $\hat{\beta}$.

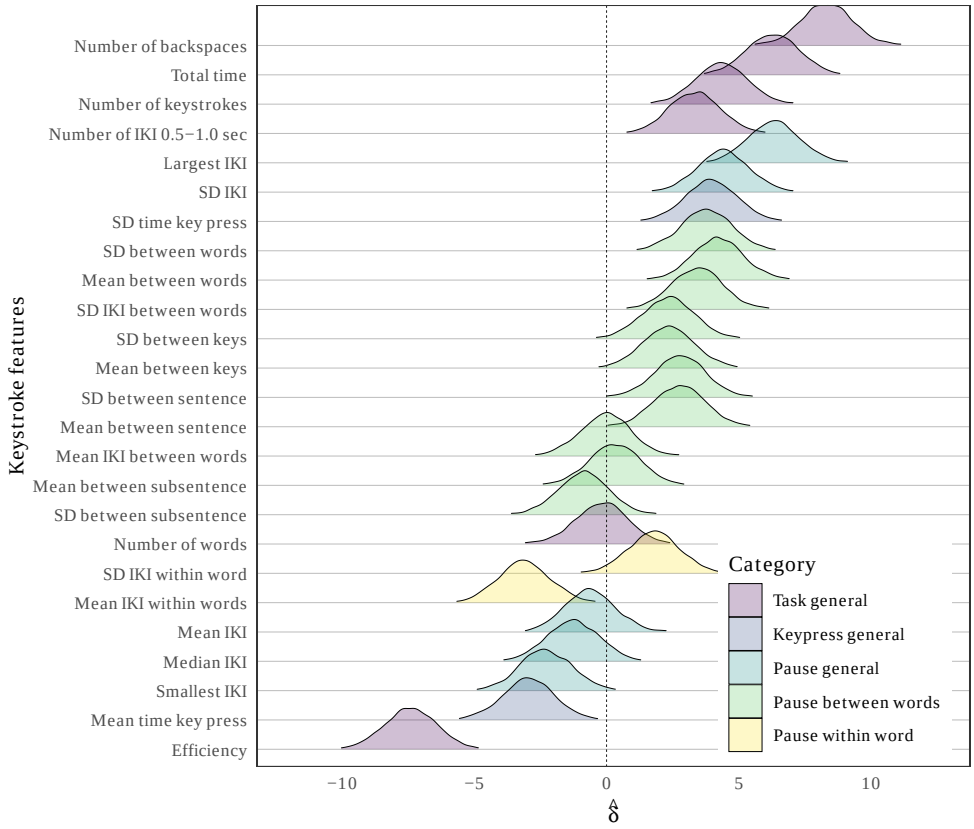

Figure 3.4: Task effect strength $\hat{\delta}$ (email writing - copy task) per keystroke feature for Villani dataset. Distributions are grouped by category and effect strength.

3.3.1 Effect strength of task

The posterior distributions of the effect strength are visualized in Figure 3.4 for the Villani dataset, and in Figure 3.5 for the academic writing dataset. The keystroke features were assigned to five groups that have been identified in previous studies: features related to the task in general, key presses, latencies or pauses in general (not location-specific), pauses within words, and pauses between words (Grabowski, 2008; Wallot & Grabowski, 2013).

For the Villani dataset, firstly, the task effect was largest for features related to the task as a whole, with the largest positive effect on the number of backspaces and total time, and the largest negative effect on the efficiency. Thus, the email writing task consisted of more

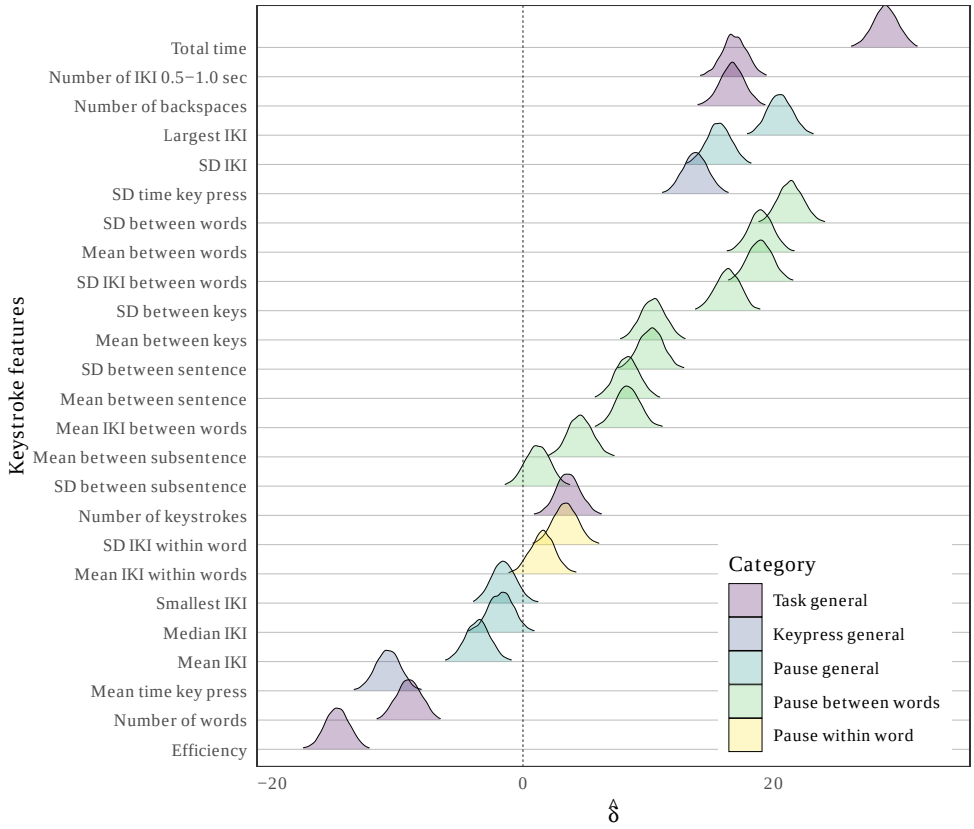

Figure 3.5: Task effect strength $\hat{\delta}$ (academic summary - copy task) per keystroke feature for academic writing dataset. Distributions are grouped by category and effect strength.

backspaces, a lower ratio of characters typed to the characters in the final product, and took longer to type, compared to the copy task. Second, features related to pausing, such as mean and median IKI, showed low effect strengths, while features related to the variance in the general pauses, such as the largest IKI and *SD* IKI, showed large effect strengths. Third, the effect on location-specific pause features was relatively low, except from the mean and the *SD* of time and the *SD* IKI between words. Fourth, the mean IKI within words showed a small negative effect, while the effect on the variance between words was positive. This indicates that participants typed faster within words, but with a larger variance, in the email writing task compared to the copy task. Lastly, the mean key press time showed a negative

effect, while the *SD* key press time showed a positive effect.

For the academic writing dataset, firstly, again the task effect was largest for features related to the task as a whole. Total time showed a large positive effect, and efficiency showed the largest negative effect. Thus, the academic writing task took longer, and had a lower ratio of characters typed to the characters in the final product, compared to the copy task. Second, for the features related to pauses in general, a large effect was found for the *SD* and largest IKI, but a small effect for the mean and median IKI. Third, the features related to location-specific pauses between words and between sentences showed relatively large effect strengths, especially for the mean and *SD* of time and the *SD* IKI between words. The effect strengths of the mean and *SD* of time between (sub)sentences were considerably smaller. Fourth, the pauses within words showed little effect. Lastly, the *SD* key press time showed a positive effect, and a slight negative effect for the mean key press time.

Conceptually related features showed similar patterns within the dataset. For example, for both mean and median IKI, the difference between tasks could be both positive and negative, rendering a small effect strength. In addition, the *SD* of time between words and the *SD* IKI between words showed a positive effect within both datasets, with similar effect sizes. However, the mean IKI between words and mean time between words did not show similar patterns in the Villani dataset: the effect of task on the mean time between words was positive with a relatively large effect strength, while the effect of task on the mean IKI between words had a small effect strength, where the direction could not be determined.

3.3.2 Differences in effect strength between datasets

The five groups of keystrokes showed similar patterns in effect strengths between the two datasets. However, the actual effect strength of task differed across datasets: in the academic writing dataset, the effect strengths were larger for almost all keystroke features, compared to the Villani dataset. These differences are shown in Figure 3.6. Comparable effects strengths across datasets (reliable effects) were found for the number of keystrokes, median IKI, smallest IKI, and *SD* IKI between words. Total time showed the largest difference in effect strength across datasets, indicating a task specific effect (email writing/academic writing), rather than a task general effect on total time. In addition, especially the variances in pause times (*SD* time between words, *SD* time between keys, *SD* IKI between words, *SD* time between sentences, *SD* IKI, and *SD* time key press) showed large differences in ef-

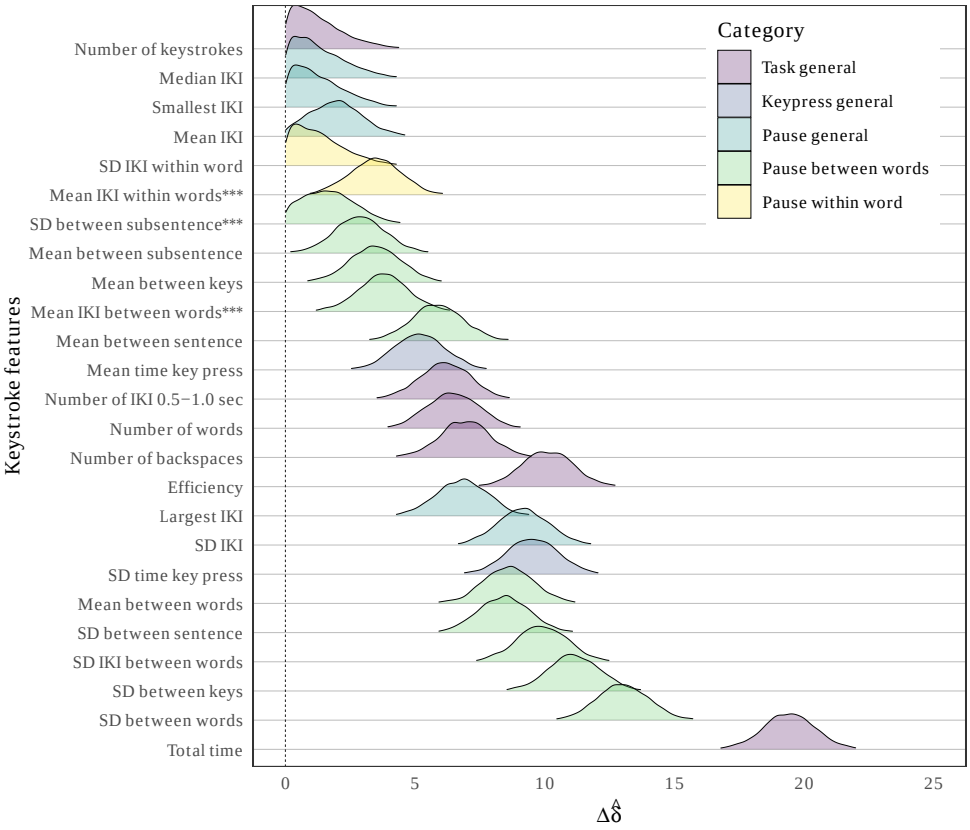

Figure 3.6: Absolute difference Δ of the effect strength $\hat{\delta}$ contrasting both datasets. Differences are shown by keystroke feature. Values close to zero indicate similar effects in both datasets. Larger values indicate effects that are different for the datasets and thus, specific to the email/academic writing task. Distributions are grouped by category and effect strength.

Note. *** Keystroke features for which the direction of effect differed between datasets.

fect. Moreover, task general effects such as efficiency, number of backspaces and number of words showed large differences in effect.

3.4 Discussion

In this chapter we aimed to investigate which, and how, keystroke features are affected by differences in cognitive demand across writing tasks. To achieve this we extracted various

keystroke features which are related to pause durations (general and location-based), and content and revising behavior. The keystrokes were compared across two different datasets, both containing a copy task and one containing an email writing task and the other an academic summary task. Bayesian linear mixed effects models were applied to determine the strength and direction of the effects of task between the different keystroke features within and across datasets. Some keystrokes showed an effect of task in both datasets, some in only one dataset, and some did not show an effect in either dataset.

First, several keystrokes features differed between the tasks in both datasets. It was hypothesized that features related to the time between words and sentences, and the amount of revisions, would differ across tasks, because these are frequently associated with cognitive processes, such as planning and revising (Van Waes et al., 2014; Wengelin, 2006). This was confirmed in both datasets. In particular, features related to the task as a whole, such as the number of keystrokes, the number of backspaces, efficiency, largest IKI, and total time, were different between the two tasks in both datasets. In addition, the mean time between words differed between writing tasks. These findings reproduced across not just the present datasets, but were also reported by other studies (Conijn & van Zaanen, 2017a; Grabowski, 2008). These features seem to be strongly influenced by the writing task, are not specific to datasets and, therefore, must be sensitive to task characteristics, such as cognitive demand. This allows the use of these features for task classification, at least for the tasks reported in those studies.

Second, some keystroke features related to time between words and sentences only showed differences in effect in the academic dataset, but not in the Villani dataset. The mean of the IKI between words, time between keys, time between subsentences, and the standard deviation of the IKI between words, time between keys, and time between subsentences, as well as the number of words only differed between the academic summary task and the copy task, but not between the copy task and the email writing task. One possible explanation for this is that these features are only affected by the task, if the difference in the cognitive demands are larger. In the present datasets, the academic summary task could be considered more complex compared to the email writing task, because it involves additional complexity, such synthesizing, integrating sources, and utilizing a repertoire of linguistic practices appropriate for the task (Lea & Street, 1998). Therefore, these features might be less sensitive to small differences in complexity or cognitive demand.

Third, it was hypothesized that keystroke duration within words would not be sensitive to task because these are associated with motor processes or individual typing skills (Grabowski, 2008). Indeed, it was shown that the mean and standard deviation of the IKI within words did not differ between tasks in the academic writing dataset. This could indicate that the cognitive writing processes during word production, beyond motor processes and typo revisions, are limited, or that cognitive writing processes within words are reflected similarly in the IKI within words in both tasks.

In addition, we found that conceptually related keystroke features, such as mean and median IKI, had similar effects within the dataset. Interestingly, the effect of task on the mean time between words and the mean IKI between words differed in the Villani dataset: the mean time between words showed a positive effect of task with a relatively large effect strength, while the mean IKI between words showed that the effect could be both positive and negative, with a really small effect strength. A possible explanation lies in the different measurements of these features. The mean time between words is the whole pause between words, while the IKI between words only measures the IKI of the last letter of the word and the 'space' key pressed. This would suggest that the feature time between words more easily picks up on the differences in task, compared to the somewhat lower-level feature IKI between words.

Extending on earlier research, we not only showed which keystroke features differed across tasks, but also compared the strength and the direction of the effects within and across datasets. For the Villani dataset, the effect of task was largest for the number of backspace keys, the largest IKI, total time, and efficiency. Thus, in the email writing task more backspaces were used, the largest IKI was longer, the total time spent was longer, and the efficiency was lower, compared to the copy task. In the academic writing dataset, the largest effects were found for total time, *SD* between words, the largest IKI, and efficiency. This indicates that students spent more time, had more variance in time between words, longer largest IKI, and lower efficiency in the academic summary task, compared to the copy task.

When comparing the effects across datasets, it was found that total time, *SD* of time between keys and *SD* of time between words are the keystroke features that differ most in the effect of task across the two datasets. Since the two datasets were assumed to vary in terms of the complexity of the writing, as opposed to the copy task, this might indicate

the usefulness of these features for determining task complexity or cognitive demand. The *SD* of time between subsentences and mean IKI between and within words even differed in direction of the effect across the datasets. The change in direction of the effect of task across datasets might indicate that these features are more related to the specific dataset rather than to the effect of task. For example, the language or style could be more complex in the Villani copy task compared to the email writing task, while the language or style in the academic dataset copy task could be less complex compared to the academic writing task.

3.4.1 Limitations and future work

The current study is limited in three ways. First, we compared two tasks in two different datasets. We argued that some of the differences in keystroke features might be due to the task complexity or cognitive demand, which differed across tasks. However, this might also be caused by other task characteristics, which we did not measure. The copy tasks were non-identical. Nevertheless, because both copy tasks did not require higher-level cognitive processes, such as linguistic planning, the differences can still be explained by the task complexity or cognitive demand. In addition, the differences might be due to other task characteristics, such as required style. However, for the purpose of this chapter we were not interested why the keystroke features differed, but merely which and how.

Second, we did not explicitly measure the complexity or the cognitive load demand of the task. Thus, we cannot specifically state the exact relation between cognitive load and the keystroke features. For example, we do not know whether the relation between the time between words and cognitive load will be linear. Although beyond the scope of the current chapter, it would be interesting for future work to further investigate the influence of cognitive load on the keystroke features. This could be done by comparing the keystroke features of multiple tasks of which the cognitive load or complexity is known, for example, by using a secondary task or questionnaire (e.g., Paas et al., 2003). This information might also be used to identify when a task is too complex or requires too much cognitive load.

Third, the differences in the keystroke features might be caused by other factors that we did not test. Keystrokes are found to be sensitive to other factors, such as handedness, keyboard type (Gunetti & Picardi, 2005; Tappert et al., 2010), typing and writing experience and abilities, environmental conditions (Gunetti & Picardi, 2005), and cognitive impairments (e.g., Van Waes et al., 2017). We do know that the participant samples differed

between the datasets (students versus students and faculty), which might indicate differences in writing experience. Yet, participant specific variation was statistically accounted for, so the differences across the dataset cannot be explained by individual differences in the samples. However, the same approach used in the current chapter could be used in future work to identify the influence of these other factors on the keystroke features. In this way, we can identify which and how keystroke features are sensitive to individual differences and experimental factors. This could indicate which factors need to be controlled for when analyzing specific keystroke features. For example, when handedness does not appear to influence the number of backspaces, handedness does not need to be controlled for when analyzing the effect of the number of backspaces between writers on the dependent variable of interest.

3.4.2 Implications for theory and educational practice

Although previous work has hypothesized that some of these features are related to cognitive demand, in this chapter we specifically showed which features varied, and how these features varied with differences in cognitive demands across tasks. These findings provide insight into which features are of interest when we are looking for evidence of cognitive writing processes, such as planning, translating, and reviewing processes (Flower & Hayes, 1980) in the keystroke log. In addition, the sensitivity of the keystroke features across tasks shows that caution should be taken when generalizing the effect of these features across tasks, because these features may differ merely as a result of the task, rather than as a result of the variable of interest, for example, writing quality, which has frequently been predicted in writing research (e.g., Allen, Jacovina, et al., 2016; Likens et al., 2017; M. Zhang et al., 2016).

Next to these theoretical implications, the findings of the current chapter have implications for educational practice. This chapter showed which keystroke features differ across tasks with different cognitive demands, and hence might be used to determine differences in cognitive load between tasks. Teachers and instructional designers can use these insights to identify differences in cognitive demands imposed by their chosen learning designs. This allows them to automatically evaluate whether their chosen writing tasks are producing the expected learning processes and outcomes (Kennedy & Judd, 2007; Lockyer et al., 2013). In addition, as keystrokes are measured during the writing process, differences in cognitive load may be determined during a single task. This can be used, for example, to deter-

mine cognitive load during different writing processes, such as planning, translating, and reviewing (cf. Alves et al., 2008). These insights can be used by teachers to determine when or with which writing processes a student needs support to improve their writing process (Santangelo et al., 2016).

3.5 Conclusion

To conclude, this chapter provides insight into how keystroke-based features differ across writing tasks with different cognitive demands. Features related to interkeystroke intervals in general, or interkeystroke intervals within words did not differ across task. Features related to the time between words or sentences, such as the *SD* of interkeystroke between sentences, or mean interkeystroke interval between words, only differed between tasks with larger differences in cognitive demands. Lastly, features related to task as a whole, such as the number of words typed, amount of revision, and total time, as well as the time between words were found to differ across all tasks. This indicates that especially these latter features are related to cognitive load or task complexity, and hence are of interest for analyzing cognitive writing processes. In addition, this chapter showed that it is important to be mindful when deriving conclusions from individual keystroke features, because they are already sensitive to small differences in writing tasks. Hence, this chapter provides us with a better understanding of the keystroke features frequently used in the writing literature.

4

Using keystroke data for early writing quality prediction

Adapted from: Conijn, R., Cook, C., van Zaanen, M., & Van Waes, L. (under review). *Early prediction of writing quality using keystroke logging.*

In the previous chapter we showed how certain keystroke features can be used to gain insight into students' writing processes. For providing automated feedback on the writing process, we also need to determine whether keystroke features can be used to identify students at risk during writing, to support students during writing. Currently, most literature focuses on providing human or machine support only after a draft is submitted. In this chapter, we investigate the use of keystroke analysis to predict writing quality already during the writing process. Keystroke data were analyzed from 126 English as a second language learners performing a timed, academic summarization task. Writing quality was measured by final grade. Based on previous literature, 54 keystroke features were extracted. Correlational analyses were conducted to identify the relation between writing process and writing quality. Next, machine learning models (regression and classification) were used to predict final grade and classify students at risk at several points during the writing process. The results show that the relation between writing quality and writing process was rather limited. In addition, the relation between the writing process features and writing quality changed over time during the writing process. At the end of the chapter, pedagogical implications are discussed for the use of keystroke analysis for providing process feedback during writing.

Acknowledgements. I would like to thank Diana Schmalzried and Xinran Wang for their assistance in collecting the academic writing dataset.

4.1 Introduction

Feedback on writing plays a key role in improving writing quality and writing proficiency (Bitchener et al., 2005; Chandler, 2003; Graham et al., 2015; Parr & Timperley, 2010). For feedback to be effective, it needs to be timely and frequent (Cotos, 2015; Ferguson, 2011). However, providing timely feedback on academic writing in higher education is complex and time-consuming. Accordingly, several automated feedback systems have been developed to augment teacher feedback and consequently also enhance students' writing proficiency (Dikli, 2006; Stevenson & Phakiti, 2014).

A key part in most of these systems is the automated scoring of students' writing, as this measure of writing quality can indicate students at risk of failure. Automated scoring systems have demonstrated to provide fairly accurate predictions of human scores, based on properties of the writing product, e.g., number of words in a draft (Allen et al., 2015). In this way, a score is predicted once a writing product (draft or final version) is finished. However, for timely feedback, we would like to be able to predict writing scores as soon as possible, in order to assist students at risk even before the writing product is finished (Romero & Ventura, 2019).

For the early prediction of writing scores, i.e., prediction of writing scores before the draft has been finished, there are two possible approaches. First, we can predict writing scores based on snapshots of the text produced so far. However, the text produced so far might not contain enough information (e.g., limited number of words) to provide accurate predictions. Second, we can predict writing scores based on information on the writing process, e.g., the number of revisions made. This approach has the additional advantage that it may provide feedback both *during* and *on* the writing process. Feedback on the writing process is more powerful for deep processing than feedback on the product, as it can not only be used to improve the current task, but also to enhance skills that can be transferred to future tasks (Hattie & Timperley, 2007; Vandermeulen et al., 2020).

Several studies already showed that information on the writing process can be used to accurately predict holistic grades (e.g., Allen, Jacovina, et al., 2016; Guo et al., 2018; Sinharay et al., 2019). In these studies, keystroke logging was used to extract information on the students' writing processes (Leijten & Van Waes, 2013; Lindgren & Sullivan, 2019). However, these studies used a varying set of keystroke features, making it hard to determine which features are most relevant for predicting writing quality. In addition, these studies

only analyzed the writing process *after* it was finished, using the full keystroke log. Hence, it is still unknown whether it is possible to use information on the writing process for the early prediction of writing scores.

Therefore, in this chapter, we aim to further explore the demonstrated relationship between writing process and writing quality by indicating which keystroke features identified in previous literature are related to writing quality. In addition, we aim to identify whether these features can be used for timely writing quality prediction at different stages in the writing process. We measured writing quality as the final score of English as a second language (ESL) students writing an academic summary. Keystroke analysis is used to automatically extract information on students' writing processes.

4.1.1 Timing of writing feedback

Almost universally, human feedback on writing is provided after the completion of a draft or final version (Gielen et al., 2010; Graham et al., 2015; Parr & Timperley, 2010). These feedback timings can be explained by the fact that the assessment and resulting feedback is commonly aimed at (near) finalized products. Teachers provide little feedback during writing, as this requires insight into the writing process, which is primarily obtained via time-intensive methods such as thinking-aloud and observations (see e.g., Beauvais et al., 2011; Braaksma et al., 2004).

While human feedback on the written product usually only occurs once or twice per writing assignment, computer-based support for writing allows for timely assessment and feedback on multiple drafts (Cotos, 2015). There is a large variety of computer-based writing support systems available to assist teachers in providing writing support, such as automated essay scoring (AES), automated writing evaluation (AWE), and intelligent tutoring systems (ITS; Allen et al., 2015). AES are grading systems typically used for summative assessment that can be used as either an alternative to teachers grading or as a first-draft evaluator (Dikli, 2006; Wilson, 2017). AWE systems are intended as formative assessment tools, providing more detailed feedback and suggestions for improvement than AES systems (Cotos, 2015). Lastly, ITS extend on AWE systems by also providing instructional content, probing questions and interactivity (Ma et al., 2014). The feedback of these systems over the course of multiple drafts can constitute feedback during the writing process to some extent. However, as in human support, the automated assessment and feedback is still commonly based on the writing product (or intermediate writing products), rather

than the writing process (Cotos, 2015; Ma et al., 2014; Wang et al., 2013). Accordingly, the feedback commonly aims at revisions at the micro (product) level, such as grammar and wording, rather than at support for the development of writing strategies and self-monitoring (Strobl et al., 2019). Therefore, in this chapter, we focus on the writing process; for an overview of automatic feedback on the writing product, see e.g., Crossley et al. (2019) and Dikli (2006).

4.1.2 Measuring writing process with keystrokes

Keystroke analysis has been increasingly used to gain insight into students' writing processes. Keystroke logging can provide objective, detailed, and real-time information on students' unfolding typing processes during their writing (Leijten & Van Waes, 2013; Lindgren & Sullivan, 2019). Given the fine-grained nature of keystroke logging data, a variety of features have been extracted for keystroke analysis. Based on previous literature, we distinguish five different groups of keystroke features: (1) features related to latencies, such as interkeystroke intervals or timings between words (Barkaoui, 2016; Medimorec & Risko, 2017); (2) features related to revisions, such as the number of backspaces or the duration of backspacing events (Barkaoui, 2016; Deane, 2014); (3) features related to verbosity, such as the number of words (Allen, Jacovina, et al., 2016; Likens et al., 2017); (4) features related to fluency, such as the percentage of bursts ending in a revision (Baaijen et al., 2012; Van Waes & Leijten, 2015); and (5) features related to events other than keystrokes producing characters, such as text selections, insertions (paste), deletions (cut), and mouse movements (Baaijen & Galbraith, 2018; Leijten, Van Waes, et al., 2019).

4.1.3 Relation between writing process and writing quality

Before writing processes were measured using keystroke logs, several relations have been found between higher-level writing processes, such as planning and revision, and writing quality. In terms of planning, preparing a written outline enhances writing quality (Kellogg, 1987). Drafting style (rough versus polished) did not have an effect on writing quality (Kellogg, 1987). However, other studies did find an effect, where detailed plans resulted in higher writing quality compared to minimal drafts (Torrance et al., 2000). In terms of revisions, more proficient writers revise more and focus more on meaning level revisions, compared to less proficient writers who focus more on surface level revisions, such as punctuation, spelling, and wording (Choi, 2007; Faigley & Witte, 1981). The influence of these

writing processes on writing quality differs over time. For example, reading the assignment and evaluating the text written so far are positively related to quality in the beginning, but negatively in the middle of the writing process. Likewise, goal setting, generating ideas, structuring, rereading, and writing are positively related to writing quality at the end of the writing process, but negatively or not related in the beginning (Breetvelt et al., 1994).

With the advent of keystroke logging in writing research, more fine-grained measures of writing processes have been related to writing quality across a variety of tasks, such as argumentative and policy recommendation essays (Guo et al., 2018; M. Zhang et al., 2016), as well as persuasive essays (Allen, Jacovina, et al., 2016; Deane, 2014; Likens et al., 2017; Sinharay et al., 2019; M. Zhang et al., 2016). First, total time on task has been shown to be correlated positively with writing scores in several studies across several tasks, with correlations ranging from .40 to .52 (Guo et al., 2018; Sinharay et al., 2019; M. Zhang et al., 2019). Features related to latencies, such as IKI within words, have been found to be negatively related with writing scores (r = -.36; Sinharay et al., 2019), while vectors of interword intervals have been found to be positively related (r = .46 to .48; M. Zhang et al., 2016). In addition, for revisions, students with low second language proficiency made more revisions, and especially more typographic, language, and pre-contextual revisions (revisions at the leading edge), compared to students with high second language proficiency (Barkaoui, 2016; Xu, 2018). Moreover, features related to verbosity, such as the number of keystrokes (r = .59; M. Zhang et al., 2019) or the number of words (r = .53; Likens et al., 2017) are positively related to writing scores. Lastly, features related to writing fluency, such as the typing speed (r = .31 to .39; Sinharay et al., 2019; M. Zhang et al., 2019), number of bursts (sequences of keystrokes without a long pause; r = .49), or burst length (r = .38) have also been found to be positively related with writing quality (Sinharay et al., 2019).

Next to the correlational analyses with single features, multiple features have been combined to predict writing quality. For example, Allen, Jacovina, et al. (2016) showed that the number of words, number of backspaces, maximum and median interkeystroke interval, as well as the standard deviation, entropy, and maximum of the number of keystrokes per 30-second interval, and the standard deviation of the distance between 30-second windows with at least one keystroke, could explain 74% of the variance in essay scores. The number of keystrokes was found most predictive of essay score. A later study using the first 999

keystrokes of the same dataset showed that 28% of the variance in essay score could be explained by the number of words (Likens et al., 2017). Furthermore, fractal properties from the multifractal analysis on the IKI timeseries combined with the number of words could explain 35% of the variance in the essay score. Using boosting with regression trees, Sinharay et al. (2019) were able to predict writing scores with 38 process features, leading to an RMSE of 0.50 (on a scale from 1–5). Time on task, typing speed, number of bursts, and burst length had the most predictive power. The prediction with product features was only slightly better (RMSE = 0.44) compared to the process features. Adding process features to the product features did not enhance the prediction accuracy.

Lastly, some studies used feature reduction on the keystroke features prior to the prediction of writing quality. For example, Deane (2014) identified three factors in the writing process: latency, editing behavior, and burst span. Combined, these factors could explain 60% of the variance in essay scores for the persuasive task, and 68% of the variance for the literacy analysis task. Another study identified two factors: sentence production and global linearity (Baaijen & Galbraith, 2018). These factors, however, did not show significant correlations with text quality.

Given the differences in sample sizes, grading procedure, writing task, writing environment, keystroke features extracted, and analyses used, the results of these studies are hard to compare. Yet, they do provide some insight into which features are related to writing quality, and to what extent writing quality can be predicted using information from the writing process. However, these studies used different and relatively small sets of keystroke features, making it hard to determine which of the features are most relevant for predicting writing quality. In addition, the keystroke features in the reported studies were analyzed after the writing process was finished. It is still unknown at what stage in the writing process keystroke data can be used for timely prediction of writing quality.

4.1.4 Timely identification of students at risk

Timely identification of students at risk is a common theme in the fields of learning analytics and educational data mining (Romero & Ventura, 2019). It has been shown that students at risk could be identified relatively quickly in a variety of contexts and with a variety of datasets. For example, on the course level, learning management system data can be used for identifying students at risk early on in the course (Macfadyen & Dawson, 2010) and historical data on grades and courses taken can be used to determine students at risk

even before the course has started (Polyzou & Karypis, 2019). On the task level, prior performance, hint usage, activity progress, and interface interaction can be used to predict successful completion of block-based programming tasks (Emerson et al., 2019) and clickstream data can be used to predict successful completion of a novice programming task already within the first minute of the task (Mao et al., 2019).

To the authors' knowledge, no studies have looked into the early prediction of students at risk using keystroke data, with the exception of Casey's (2017) study using keystroke data to predict performance in a programming course after every week in the semester. However, no keystroke studies looked into early identification of students at risk during academic writing.

4.1.5 Current approach

In the current chapter, we aim to determine the relation between the writing process (measured by keystroke data) and writing quality (measured by final grade). For this, three analyses were conducted. First, correlational analyses are used to determine which keystroke features (obtained from previous studies) are related to writing quality. Second, machine learning algorithms are trained to predict writing scores (regression). Third, machine learning algorithms are trained to predict students at risk (binary classification) at different stages in the writing process. These predictions are used to determine which keystroke features can be used for the prediction of writing quality and how prediction accuracy changes over time.

4.2 Method

4.2.1 Participants

The data used in this study were collected during an academic writing course for ESL learners taught for premaster students of communication and information sciences. The study consisted of an online part and a part during the lecture. In total, 141 students participated in the online part, 131 students participated in the lecture (this formed the academic writing dataset described in Chapter 3), and 130 students participated in both parts. In this chapter we only report on the students who participated in both parts. Of the 130 participants, 87 (67%) were female and the average age was 24 (SD = 2.8).

4.2.2 Procedure

In the online part, the participants were asked to provide informed consent and to complete a questionnaire on demographics and self-reported writing style. In addition, they were asked to read a given journal article in preparation for the lecture the following week. The article involved a 2 x 2 experimental design setup in the field of their premaster program (Woong Yun & Park, 2011). A week after the questionnaire, the participants were assigned two writing tasks during the lecture. The first task was a copy task in which students were asked to transcribe a given fable of 850 characters. The second task was an academic summary task, where the participants were asked to write an academic summary of 100–200 words based on the article they read in the week prior to the lecture (the abstract was removed from the article). The participants were allowed 30 minutes to finish this task. Five minutes before the end of the task, they were reminded to finish their writing.

All students used similar desktop computers for the task. The task description was shown on a single page at the left of the screen with the Word document where the participants could type the summary on the right. The text of the journal article was added underneath the task description in the same document. To consult the journal article, participants had to focus (click) on the task description and scroll down. The participants were allowed to use the Internet during the task (for example, to consult an online dictionary). During the tasks, keystroke data and mouse data were collected using Inputlog (Leijten & Van Waes, 2013).

4.2.3 Data collection

For the current study, the keystroke data of the academic writing task and the essay scores were analyzed. The keystroke data of the copy task were used as a baseline task for extracting the keystroke features. Of the 130 participants, two participants did not type in the specified Word document, one participant only completed the copy task, and one participant only copy-pasted text. Therefore, these participants were excluded, resulting in a total of 126 participants remaining for analysis.

The academic writing tasks were independently graded by a native English speaker and an ESL speaker, both highly experienced in grading writing. The writing tasks were scored against five rubrics (see Appendix A): main idea, structure and organization, content, language and paraphrasing, and grammar and mechanics from 1 (not passing) to 4 (exceptional). The final grade was calculated by the sum of the rubric grades divided by 2. Points

Table 4.1: Descriptive statistics and inter-rater reliability of final grade and rubric scores (N = 126)

Rubric	Mean grade	*SD* grade	N at risk	Inter-rater reliability
Main idea	2.41	0.74	65 (52%)	κ = .62
Structure and organization	2.61	0.69	54 (43%)	κ = .51
Content	2.02	0.76	101 (80%)	κ = .63
Language and paraphrasing	2.75	0.82	31 (25%)	κ = .76
Grammar and mechanics	1.87	0.93	91 (72%)	κ = .86
Final grade	5.50	1.36	60 (48%)	r = .88[a]

Note. κ = quadratically weighted kappa, r = Pearson's correlation. At risk is $\leq$ 2 for the rubric scales, and $<$ 5.5 for the final grade. [a] Correlation was chosen over quadratically weighted kappa for the final grades (κ = .89) The final grade includes half points, and hence could result in inter-rater differences of 0.5, which are beneficially penalized in the quadratic weighing (0.5^2 = 0.25). Therefore, the quadratically weighted kappa would be overly optimistic. In addition, final grade is measured on a different scale than the rubric scores, which could result in faulty conclusions drawn when comparing the rubric scores' inter-rater reliability with the reliability of the final grade.

were subtracted if the student did not comply with the task (e.g., wrote too few or too many words, or did not cite the authors). This resulted in a scale from 1 to 10. A student was considered at risk if the final grade was lower than 5.5; a final grade of 5.5 or higher was considered a pass. Inter-rater reliability was calculated for all scores using a quadratically weighted kappa to account for the ordinal scale (as in M. Zhang et al., 2019). As in M. Zhang et al. (2019), the grades from the first grader (native English speaker), were used for analysis. The descriptive statistics and inter-rater reliability of the grades are shown in Table 4.1.

4.2.4 Feature extraction

Based on existing literature, a total of 54 features were extracted from the keystroke log. The literature sometimes uses different definitions for similar keystroke features and it is sometimes unclear how exactly a certain feature is extracted. For replicability, we provide a detailed overview of the extraction of features from the keystroke log. To be able to replicate the features from previous studies as closely as possible, we did not use the analysis tool provided by Inputlog, but used the raw data ('basic log file') to extract the features in R. The R scripts for the feature extraction, feature reduction, and model building can be accessed at `https://github.com/RConijn/Early-Prediction`.

Inputlog collects both keystroke and mouse data and distinguishes five types of events: keyboard, mouse, insert (insertion of text from within the document or other source), replacement (selection of text), and focus (click on another window, e.g., another document or web page). Here, we define a keystroke as a keyboard event. This includes any key pressed and includes character keys (e.g., 'a', '6', or '$') as well as control, function, or navigation keys (e.g., 'Alt', 'F5', 'Home'). Sometimes multiple keystrokes are required to generate one character (e.g., for capitalization). For every character typed, the location in the document where it is typed is stored. We distinguished two locations: leading edge (at most two characters away from the end of the text), or somewhere else in the text (cf. pre-contextual versus contextual revisions in Lindgren & Sullivan, 2006a).

The features identified from previous studies can be categorized into features related to latencies, revisions, typing bursts, verbosity, and other (non-typing) events. An overview of the features and their descriptive statistics are presented in Table 4.2.

Features related to latencies. The majority of features we extracted are related to latencies. Most features are related to the interkeystroke interval (IKI), the time from a key press until the next key press. All time-based features showed a large positive skew. Therefore, these features (except for total time, initial time, and maximum IKI) were log transformed and all values above the 95th percentile were removed. Similar approaches can be found in previous studies (e.g., Grabowski, 2008; Van Waes et al., 2017).

- **Initial pause time**. Time from assignment start until first key press (Allen, Jacovina, et al., 2016; Sinharay et al., 2019).
- **Total time**. Time from assignment start to last key release (maximum is 30 minutes; Allen, Jacovina, et al., 2016; Deane, 2014; Guo et al., 2018).
- **Mean, Median, *SD*, and maximum IKI**. Metrics of the time from a key press until the next key press (Allen, Jacovina, et al., 2016; Sinharay et al., 2019).
- **Mean and *SD* IKI within word**. Metrics of all IKI of keystrokes within words (Deane, 2014; Sinharay et al., 2019).
- **Mean and *SD* IKI between words**. Metrics of all IKI of keystrokes between words (Deane, 2014; Sinharay et al., 2019).
- **Mean and *SD* time between words**. Metrics of the time from key press of the last letter of a word until the key press of the first letter of the next word (Deane, 2014; Guo et al., 2018; M. Zhang et al., 2016).

Table 4.2: Descriptive statistics and correlational analyses of the keystroke features over the complete writing process and begin, middle, and end of the writing process

		Correlation (*r*) final grade			
Keystroke feature	Mean (*SD*)	[0–30]	[0–10]	[10–20]	[20–30]
Initial pause time (min)	1.3 (1.5)	-.16	-.16	-.03	-.09
Total time (min)	27.4 (3.6)	-.01	.02	.06	-.03
Mean IKI	174.7 (24.2)	-.10	-.12	.01	-.09
SD IKI	133.9 (15.7)	.14	.24**	.03	.18
Median IKI	139.8 (22.4)	-.06	-.04	.03	-.05
Largest IKI (min)	0.44 (0.20)	-.10	-.02	.00	-.04
Mean IKI within word	146.7 (21.6)	-.02	.04	.01	-.04
SD IKI within word	68.4 (7.9)	.02	.03	.02	-.01
Mean IKI between words	236.6 (54.4)	.03	.02	.02	.04
SD IKI between words	239.8 (42.6)	-.11	.01	-.06	-.03
Mean time between words	718 (205)	.02	.03	.05	-.06
SD time between words	894 (172)	.02	-.03	.08	-.01
Mean time between sentences	2954 (3279)	-.03	.03	-.10	-.11
SD time between sentences	4195 (3850)	.03	.01	-.20*	-.06
Number of IKI 0.5–1 s	134 (48.9)	-.09	-.02	-.11	-.05
Number of IKI 1–1.5 s	32.5 (13.2)	-.09	-.03	-.11	-.04
Number of IKI 1.5–2 s	15.2 (7.1)	-.12	-.14	-.06	-.05
Number of IKI 2–3 s	15.7 (7.5)	-.12	-.16	-.05	-.02
Number of IKI larger than 3 s	28.5 (13.4)	.04	.01	.03	.04
Perc. of long pauses between words	25% (9%)	.22*	.14	.20*	.23*
Number of revisions	89.7 (40.5)	.03	.04	.03	-.01
Number of leading-edge revisions	47.4 (46.1)	.06	.05	.07	.03
Number of in-text revisions	42.4 (29.1)	-.07	-.02	-.07	-.06
Number of backspaces	342 (173)	-.10	-.07	-.10	-.03
Mean time in single backspacing	82.4 (18.5)	.15	.12	.17	.12
SD time in single backspacing	21.7 (11.9)	-.01	.04	.05	.00
Mean time in multiple backspacing	1598 (1046)	.09	.05	.09	.06
SD time in multiple backspacing	3672 (3333)	-.02	-.10	-.01	-.16
Perc. of characters final text	54% (18%)	-.11	-.05	.03	-.04

Table 4.2: Descriptive statistics and correlational analyses of the keystroke features (continued)

		Correlation (*r*) final grade			
Keystroke feature	Mean (*SD*)	[0–30]	[0–10]	[10–20]	[20–30]
Perc. of characters at leading edge	51% (35%)	.11	.06	.07	-.02
Mean #keystrokes per burst	15.2 (5.2)	-.11	-.06	-.14	-.11
SD #keystrokes per burst	18.9 (8.5)	-.10	-.03	-.15	-.07
Largest #keystrokes per burst	115 (72.3)	-.06	-.02	-.10	-.04
Number of bursts	159 (56.3)	.02	.08	-.02	-.01
Percentage of R-bursts	7% (6%)	.12	.13	.10	.03
Percentage of I-bursts	7% (7%)	-.08	-.09	-.07	.01
Percentage of words in P-bursts	34% (22%)	.04	.00	.00	-.03
Number of production cycles	0.3 (0.2)	.15	-.01	.15	.18*
Perc. of linear transitions sentences	73% (8%)	-.13	-.08	-.18*	-.24**
Perc. of linear transitions words	11% (12%)	.03	-.01	.12	-.08
Number of keystrokes	2387 (811)	-.04	.03	-.09	-.02
Number of words	295 (97.1)	-.05	.02	-.09	-.03
SD #keystrokes per 30 s	41.1 (10.7)	-.03	.04	-.10	-.04
Slope #keystrokes per 30 s	0.0 (0.8)	-.04	-.03	-.09	-.02
Entropy #keystrokes per 30 s	0.0 (0.0)	.03	.05	.13	.05
Uniformity #keystrokes per 30 s	576 (168)	-.07	.03	-.11	-.04
Local extreme #keystrokes per 30 s	47.0 (7.4)	.12	.10	.11	.04
Mean distance 30 s window >1 key	1.3 (0.2)	-.07	-.03	.01	-.03
SD distance 30 s window >1 key	0.9 (0.7)	-.10	-.04	-.09	-.15
Number of shifts to translation	0.5 (2.7)	.08	.19*	.01	.06
Number of shifts to task	19.8 (6.6)	.13	.22*	.01	.01
Mean time cut/paste/jump events	545 (395)	.07	.01	.12	.09
SD time cut/paste/jump events	1323 (1282)	.11	.09	.17	.18*
Perc. of time spent on other events	50% (16%)	.12	.03	.12	.10

Note. All time-based features are in milliseconds (except stated otherwise); * $p < .05$, ** $p < .01$. Given the multiplicity, *p*-values should be interpreted as exploratory rather than confirmatory.

- **Mean and *SD* time between sentences**. Metrics of the time from key press of the end of a sentence marker until the key press of the first letter of the next sentence (Baaijen & Galbraith, 2018; Deane, 2014).
- **Number of IKI of specific length**. Five features were extracted: the number of IKI between 0.5–1.0 seconds, 1.0–1.5 seconds, 1.5–2.0 seconds, 2.0–3.0 seconds, and larger than 3.0 seconds (Allen, Jacovina, et al., 2016).
- **Percentage long pauses between words**. Number of pauses between words longer than two *SD* from the mean IKI within the copy task, divided by the total number of pauses between words (Baaijen & Galbraith, 2018).

Features related to revisions. Eight features related to revisions were extracted. Again, all time-based features were log transformed and all values above the 95th percentile were removed.

- **Number of revisions**. Number of insertions away from the leading edge plus the number of sequences of backspaces and delete keystrokes, that do not contain a pause longer than two *SD* from the mean IKI within the copy task, and where the cursor was not replaced to a different location in the text during the revision (Barkaoui, 2016).
- **Number of leading-edge revisions**. Number of revisions at the leading edge (pre-contextual revisions; Barkaoui, 2016).
- **Number of in-text revisions**. Number of revisions away from the leading edge (contextual revisions; Barkaoui, 2016).
- **Number of backspaces**. Number of backspaces and delete keystrokes (Allen, Jacovina, et al., 2016).
- **Mean and *SD* time in single backspacing**. Metrics of the time of a sequence of backspaces or delete keystrokes which included only one backspace or delete keystroke (Deane, 2014).
- **Mean and *SD* time in multiple backspacing**. Metrics of the time of a sequence of backspaces or delete keystrokes which included more than one backspace or delete keystroke (Deane, 2014).
- **Percentage of characters in final text**. The number of characters in the full text, divided by the total number of keystrokes (Baaijen & Galbraith, 2018).

- **Percentage of characters at leading edge**. The number of characters typed at the leading edge of the text, divided by the total number of keystrokes, used as a proxy for the size of contextual revisions (cf. Barkaoui, 2016).

Features related to fluency. Fluency in writing has been argued to be reflected in verbosity (production), process variance, revision, and pausing behavior (Van Waes & Leijten, 2015). Here, we solely focus on the burstiness of the writing. Sentences are argued to be composed in sentence parts, also known as written language bursts (Kaufer et al., 1986). Written language bursts, hereafter bursts, are defined as sequences of text production without a long pause. To account for individual differences in typing speed, all bursts are defined as sequences of keystrokes that do not contain pauses longer than two *SD* from the mean IKI within the copy task of the same participant (as in Deane, 2014).

- **Mean, *SD*, and maximum number of characters per burst**. (Sinharay et al., 2019).
- **Number of bursts**. (Sinharay et al., 2019).
- **Percentage of R-bursts**. Number of revision bursts at the leading edge ending in a revision, divided by the total number of bursts (Baaijen et al., 2012).
- **Percentage of I-bursts**. Number of insertion bursts produced away from the leading edge, divided by the total number of bursts (Baaijen et al., 2012).
- **Percentage of words in P-bursts**. Number of words in 'clean' production bursts both initiated and terminated by a long pause (not a revision), divided by the total number of words (Baaijen et al., 2012).
- **Number of production cycles**. Number of groups of bursts without interruptions by other events (i.e., all events not resulting in a character being typed, see also features related to other events), divided by the number of words (Baaijen & Galbraith, 2018).
- **Percentage of linear transitions between words**. Number of times the transition to the next word was not interrupted by other events divided by the total number of transitions between words (Baaijen & Galbraith, 2018).
- **Percentage of linear transitions between sentences**. Number of times the transition to the next sentence was not interrupted by other events, divided by the total number of transitions between sentences (Baaijen & Galbraith, 2018).

Features related to verbosity. Two general features related to verbosity were extracted. In addition, as in Allen, Jacovina, et al. (2016), we extracted seven features related to the variability of the keystrokes over time. These features are all related to the number of keystrokes in 30-second time windows. Since the task duration was 30 minutes, there were a total of 60 time windows.

- **Total number of keystrokes** (Allen, Jacovina, et al., 2016).
- **Total number of words** (Likens et al., 2017).
- ***SD* number of keystrokes per 30s**. Variance of the number of keystrokes in every 30 seconds (Allen, Jacovina, et al., 2016).
- **Slope of the number of keystrokes per 30s**. The slope of the linear regression applied to the sequence of keystrokes in every 30 second window (Allen, Jacovina, et al., 2016).
- **Entropy of the number of keystrokes per 30s**. Maximum likelihood estimation of Shannon entropy for the number of keystrokes in every 30 second window, divided by the total number of keystrokes (Allen, Jacovina, et al., 2016). Calculated with the "entropy" function in the R-package "entropy" (Hausser & Strimmer, 2014).
- **Uniformity of the number of keystrokes per 30s**. Jensen-Shannon Divergence of a uniform distribution of keystrokes (every window: total number of keystrokes divided by total number of windows) and the actual distribution of keystrokes per 30 seconds (Allen, Jacovina, et al., 2016). Calculated with the 'JSD' function in the R-package "philentropy" (Drost, 2018).
- **Local extreme number of keystrokes per 30s**. Number of times the sign of the difference in the number of keystrokes between 30 second window changes, from increasing to decreasing number of keystrokes or vice versa (Allen, Jacovina, et al., 2016).
- **Mean and *SD* distance 30s windows of more than one keystroke**. Metrics of the distance between 30 second windows with more than one keystroke, gives a measure of the amount and variance of long pauses (Allen, Jacovina, et al., 2016).

Features related to other events. Lastly, we also included five keystroke features related to non-typing or 'other' events to get a broader view of students' writing behavior. Other events are all events which do not result in a character being typed, including mouse events,

insertions, replacements, focus events, control, function, and navigation keys (Baaijen & Galbraith, 2018). Since typos are extremely common, these were not counted as an 'other' event. We operationalized a typo as a revision within a word, consisting of a maximum of three backspace or delete keystrokes and where the IKIs of the delete or backspace keystrokes are shorter than two *SD* from the mean IKI in the copy tasks.

- **Number of focus shifts to translation or task**. Since focus shifts other than to task or translation (e.g., to social media websites) were rare, we only included those two types of focus shifts: the number of times the focus shifted towards an online dictionary or translation web page and the number of times the focus shifted towards the task (Leijten, Van Horenbeeck, & Van Waes, 2019).
- **Mean and *SD* cut/paste/jump events**. Metrics of the time spent on cut (selection followed by a keystroke or insertion), paste (insertion), and jump (mouse click resulting in change in position in the document) events (Deane, 2014).
- **Percentage of time spent on other events**. Time spent on other events, divided by the total time spent (Baaijen & Galbraith, 2018).

To be able to perform early prediction, all features were calculated for the keystrokes up to different timepoints in the writing process. First, the keystroke log was divided into six equal time-based segments. Because the writing task lasted 30 minutes, this resulted in six segments of five minutes each. All 54 features were calculated for each segment (resulting in six feature sets) as well as up to each segment (resulting in five feature sets). For each model, the feature sets available up to that timepoint were included. For example, for the model at 15 minutes, the feature sets from 0–5, 5–10, 10–15, and 0–15 minutes were included.

The data indicated that four students did not start typing in the first five minutes. For these students, there is no information in the features for the prediction models at five minutes. Therefore, we also determined the performance after a certain number of keystrokes. The keystroke log was divided into six equal segments, thus each segment contained 1/6th of all keystrokes. Similar to the time-based segments, all 54 features were calculated for each keystroke-based segment and up to each keystroke-based segment.

4.2.5 Feature reduction

Given the large number of features (at least 54 per model) and the limited number of observations (126 participants), feature reduction was conducted. This was done for each model separately, as we expected that the prediction power for the features would differ across the

models; that is, certain features might work better at different times in the writing process. Thus, the final sets of features used in the prediction differed for each model. However, as the feature reduction approach was identical for all models, model comparisons are still valid.

Two filter and one wrapper method from the R-package "caret" (Kuhn, 2019) were applied. First, for each set of features collected at each segment, the features with near zero variance were removed ('nearZeroVar' function). Near zero variance features were defined as the features with a ratio of the most common value to the second most common value smaller than 95/5 or less than 10% unique values. Second, highly correlated features, with pair-wise correlations above .80 were identified ('findCorrelation' function). From these pairs of highly correlated features, the feature with the largest mean absolute correlation was removed. Lastly, recursive feature elimination was applied ('rfeControl' function). With recursive feature elimination, all features are first used to fit the model and the features are ranked according to their importance. At every next step, the model is fitted again with all features except for the predictor that had the lowest importance (according to the previous step). This process is repeated recursively. Here, the best subset of features was determined using 10-fold cross-validation, by selecting the subset of features that resulted in the lowest root mean squared error (RMSE) for the regression models, and in the largest AUC for the classification models. In addition, to avoid overfitting, a simpler model (fewer features) was preferred over a more complex model if the difference in RMSE was less than 1% ('pickSizeTolerance' function).

4.2.6 Modeling

Three analyses were conducted on the keystroke features. Firstly, correlational analyses were run to determine the relationship of the keystroke features with writing quality, measured by final grade. These correlational analyses were both conducted over the full writing process, as well as for three different segments within the writing process (0–10 minutes, 10–20 minutes, and 20–30 minutes).

Secondly, regression models were trained on the keystroke features to predict final grade at different timepoints in the writing process. Six timepoints were based on the elapsed time (every five minutes within the writing process), and six timepoints were based on the number of keystrokes (every 1/6th of the total number of keystrokes). Three regression models were run at each of those timepoints: random forest, support vector machines

with radial kernel, and naive Bayes. These three types were chosen as they generally work well on continuous data. All models were run using 10-fold cross-validation and the mean final grade was used as a baseline. Root mean squared error (RMSE) was used as an evaluation metric. We favored RMSE over mean absolute error (MAE), as we want to assign a larger penalty to larger errors.

Thirdly, binary classification models were trained on the keystroke features to predict students at risk at different points in the writing process. Participants with a score lower than 5.5 (on a scale from 1–10) were classified as 'at risk' and those with a score equal to or higher than 5.5 were classified as 'no risk'. Three classification models—random forest, support vector machines with radial kernel, and naive Bayes—were run for the same timepoints as used in the regression models (six timepoints based on elapsed time and six timepoints based on the number of keystrokes). All models were run using 10-fold cross-validation and the majority class was used as a baseline. AUC, precision, recall, and F-score were used as evaluation metrics. Lastly, the five most important features across all resamples in the recursive feature elimination are reported, to indicate which features have the highest predictive value and whether this differs over time during the writing process.

4.3 Results

4.3.1 Correlational analyses

Table 4.2 (p. 88) presents the correlations of all keystroke features with final grade. For the keystrokes measured over the full writing processes, only one significant correlation was found. Final grade had a small positive relation with the percentage of long pauses between words (r = .22). Thus, more pauses between words are related to a higher writing quality. The correlations over the beginning [0–10 minutes], middle [10–20 minutes], and end [20–30 minutes] of the writing process showed different patterns. The *SD* of the IKI (r = .24) and the number of focus shifts to translation (r = .19) and task (r = .22) were significantly related with final grade only in the beginning of the writing process. By contrast, the *SD* of the time between sentences (r = -.20), percentage of long pauses between words (r = .20 and r = .23), number of production cycles (r = .18), percentage of linear transitions between sentences (r = -.18 and r = -.24), and the *SD* of the time in cut, paste, and jump events (r = .18) were only significantly related to final grade in the middle or end of the writing process. This shows that the relation of the features with final grade differs over time within the writing process.

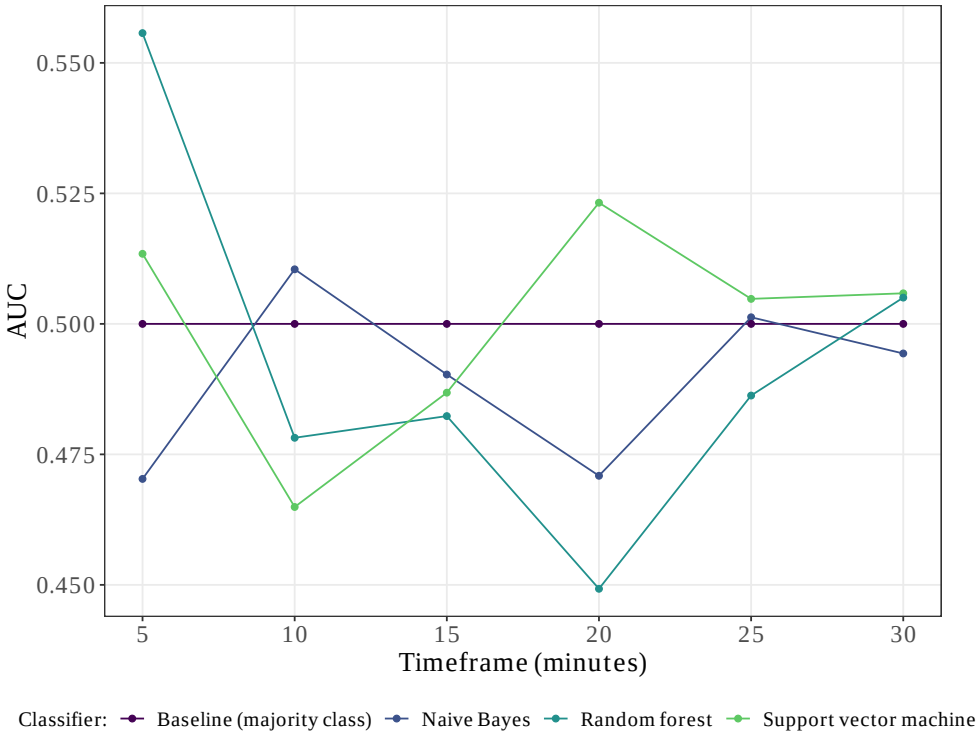

Figure 4.1: AUC of the classification models predicting students at risk, compared to the baseline, for the keystroke data up to every five minutes.

4.3.2 Predicting writing quality (regression)

First, we predicted final grade at the different timepoints (based on time elapsed and number of keystrokes) in the writing process. None of the models outperformed the baseline (mean final grade) at any of the timepoints in the writing process. Thus, the keystroke features cannot be used for the early prediction of final grade, nor for the prediction of final grade once the writing process is finished.

4.3.3 Predicting writing quality (classification)

Fortunately, for identifying students at risk, we would not need such a specific model. Rather than predicting final grade, it would already be enough to predict whether the student would pass or fail the assignment. Figure 4.1 shows the performance of the classification models predicting students at risk (final grade < 5.5) after every five minutes

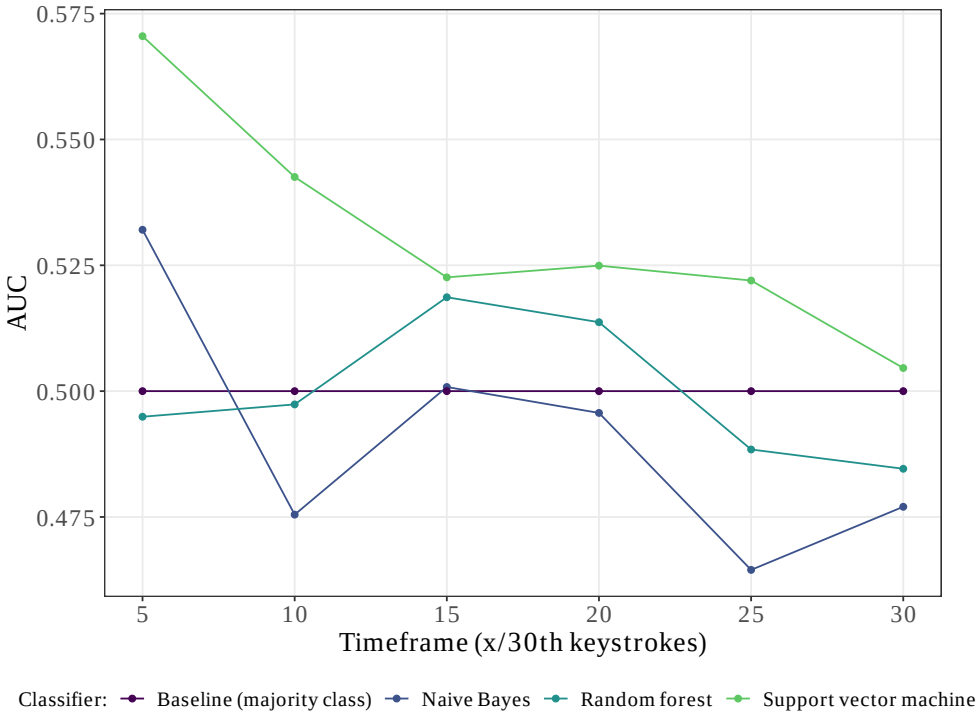

Figure 4.2: AUC of the classification models predicting students at risk, compared to the baseline, for the keystroke data up to every x/30th of the total number of keystrokes.

of keystroke data. First, it shows that the models only occasionally outperform the majority class baseline. We expected that the error would decrease when more information becomes available, in other words, when more minutes of keystroke data are used in the model. However, this trend was not clearly visible. Lastly, there does not seem to be one classification model that outperforms the other models at all points in time during the writing process.

Figure 4.2 shows the performance of the classification models predicting students at risk after every 1/6th of the total number of keystrokes. These models appear to perform slightly better than the models for every five minutes. The support vector machine outperforms the baseline and all the other models at every timepoint within the writing process. Hence, the support vector machine appears to be the best model here. However, with an AUC of 0.57, these models still do not perform well. Interestingly, the models also do

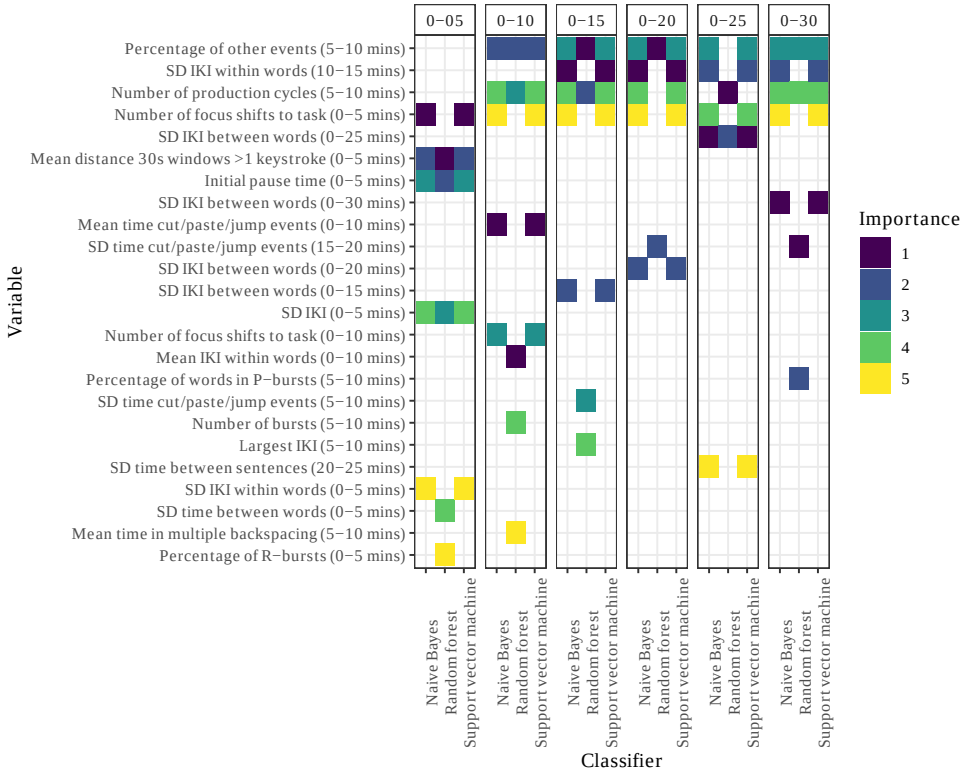

Figure 4.3: Feature importance for each model predicting at risk versus not at risk, for the keystroke data up to every five minutes.

Note. Features are ordered by importance; 1 is the best feature, 2 is the second-best feature, etc., across all resamples in the recursive feature elimination. Only the five best features per model (or fewer if the best model consisted of fewer features) are listed.

not seem to improve over time. For the naive Bayes and random forest, there seems to be an increase from 10/30th to 15/30th of the total amount of keystrokes, but when more keystrokes are added, the performance decreases again. Lastly, for the support vector machine and naive Bayes classifier, the AUC is highest when only 5/30th of the total amount of keystrokes are included in the model. This would indicate that the other keystrokes do not add much additional information for the prediction of being at risk or not.

We calculated the feature importance for all classification models to get insight into which features had the highest predictive value and whether the most important features

differed at the different timepoints. The feature importance for each model (up to the five best features) for every five minutes is shown in Figure 4.3. The feature importance was relatively consistent over the different models. This indicates that even though the models were not very accurate, there are some features that show an effect regardless of the classifier. Thus, there is still some predictive power within the features. In addition, the figure shows that the feature importance differs over time.

For the first five minutes of keystroke data, the number of focus shifts to task, the distance between 30 second windows with more than one keystroke (variance in long pauses), and the initial pause time were the most important features. When more data were taken into account (from 10 up to 30 minutes), these features were less important. Only the number of focus shifts to task in the first five minutes were still somewhat important in the later models. In the middle of the writing process, the percentage of other events (5–10 minutes), the *SD* of the IKI within words (10–15 minutes), and the number of production cycles (5–10 minutes) were most important. At the end of the writing process (20–30 minutes), no features seemed of specific importance. The *SD* of the IKI between words appeared to be most important when measured over the largest possible timeframe (e.g., 0–25 and 0–30 minutes).

The feature importance of the five best features for each model for the keystroke data up to every x/30th of the total number of keystrokes is shown in Figure 4.4. For most of the models, the mean and *SD* of the IKI of the first 1/6th of the total number of keystrokes were most important. In addition, features related to revisions in the beginning and the middle of keystroke production were found important, such as the percentage of revision bursts (10–15/30th of the keystrokes), mean time in multiple backspacing (0–5/30th of the keystrokes), and mean time in cut/paste/jump events (5–10/30th of keystrokes). No features seemed to be consistently important in the last segments of keystrokes (15–30/30th of keystrokes). In addition, no features measured over the full writing process seemed to be of specific importance.

4.4 Discussion

In this chapter, we aimed to identify how the writing process is related to writing quality and whether information on the writing process can be used for timely writing quality prediction, at different stages in the writing process. This, in turn, can be used to provide timely feedback or interventions for students at risk.

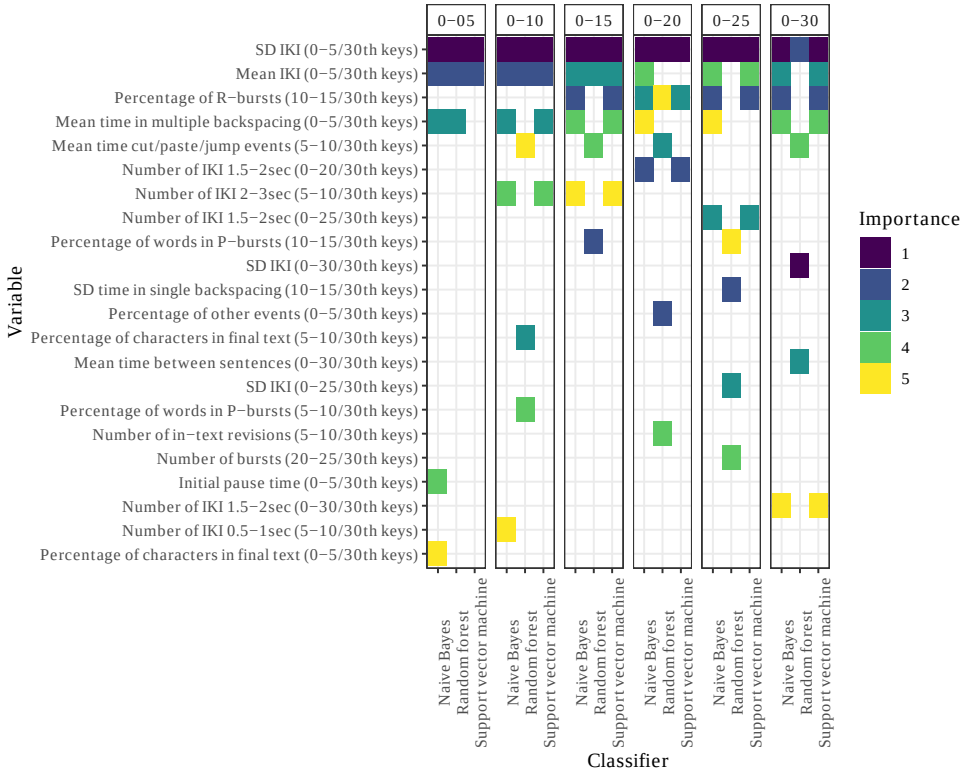

Figure 4.4: Feature importance for each model predicting at risk versus not at risk, for the keystroke data up to every x/30th of the total number of keystrokes.

Note. Features are ordered by importance; 1 is the best feature, 2 is the second-best feature, etc., across all resamples in the recursive feature elimination. Only the five best features per model (or fewer if the best model consisted of fewer features) are listed.

4.4.1 Relation between writing process and writing quality

Based on previous research on writing quality prediction using writing process features, 54 keystroke features were identified. First, correlational analyses were conducted to identify the relation between these features and writing quality. The correlation analyses showed that only the percentage of long pauses between words, measured over the full writing process, was significantly related to final grade: more pauses between words resulted in higher grades. This is in contrast with previous studies, which showed that each of the 54 keystroke features at some point have been related to final grade. For example, Allen, Jacov-

ina, et al. (2016) found medium to high effect sizes for the correlations between essay scores and the number of words, median IKI, entropy, and extremes of the number of keystrokes per 30 seconds. Sinharay et al. (2019) also found medium correlations with time within words and number of bursts. None of these were found here.

The differences in correlations can be explained by the different tasks used in these studies. Guo et al. (2018) showed that these correlations already differed for similar tasks with different prompts. Here, the differences between tasks were even larger; the current study analyzed a summary writing task, while previous studies analyzed argumentative and persuasive essay writing tasks (see e.g., Sinharay et al., 2019; M. Zhang et al., 2016). These task differences could have affected the predictive power of the keystroke features.

For example, previous work showed that especially features related to verbosity, such as total time, number of words, and number of keystrokes, resulted in moderate to high positive correlations with writing quality (see e.g., Allen, Jacovina, et al., 2016; Sinharay et al., 2019). These features were not found to be significantly related in the current study. This can be explained by differences in the task requirements. As opposed to previous studies, the present task requested students to write within a specified word limit, which a majority of the students did (only 21% of the students wrote more than 10% over the word limit). Accordingly, writing more did usually not result in higher grades; it showed that the student did not comply with the requirements, and often resulted in irrelevant information added, usually resulting in lower grades. Additionally, some previous studies used a non-timed essay task (e.g., Allen, Jacovina, et al., 2016; Guo et al., 2018), while in our study, participants were asked to write the academic summary in 30 minutes. Therefore, there is presumably less variance in some of the features, such as total time, number of words, and number of revisions, resulting in a lower predictive power of these features. This result also suggests that relative features need to be used (e.g., number of revisions per word) to avoid task length effects.

After the correlational analysis, regression analyses were conducted to model the relationship between the keystroke features and final grade. Not surprisingly, the low correlations also resulted in low prediction accuracies. Although previous studies were able to predict final grade to a large extent (e.g., Allen, Jacovina, et al., 2016; Sinharay et al., 2019), none of our models were able to predict final grade with a higher performance than the baseline. However, for the timely identification of students at risk, we do not need to

predict final grade; it might already be enough to identify students at risk of failing the assignment. Therefore, binary classification was used to determine whether the student was at risk of failing the assignment, at different timepoints in the writing process.

4.4.2 Early prediction of writing quality

For the early prediction of students at risk, keystroke data were included up to different timepoints within the writing process. Two approaches were used to identify these different timepoints based on time elapsed and the number of keystrokes. However, the accuracy was low in all cases: the models only slightly outperformed the baseline at some of the timepoints. Thus, these features do not allow for the early identification of students at risk. In addition, although the performance was expected to improve over time when more data is added, this was not found. This might be because the relation with final grade at the end of the writing process was already limited, providing little room for the performance to increase over time.

The models using the timepoints based on the number of keystrokes resulted in the best performance, with the support vector machine model consistently outperforming the baseline. This might be because by dividing the process according to the number of keystrokes, there is always information within these segments. In contrast, when dividing the process by the time elapsed, there might be limited information in the first and last segment, because the writer did not yet start or already finished their writing, respectively. Thus, although many approaches currently divide the writing process based on time (e.g., M. Zhang et al., 2016), the amount of information available in these segments should be taken into account as well.

4.4.3 Limitations

This study is limited in some ways. First, we included keystroke features which were previously found important for the prediction of writing quality once the writing process has finished. However, we showed that the importance of these features for writing quality prediction differed over time in the writing process. Hence, these features might not be the best features for the early prediction of writing quality. Future work should concentrate on identifying specific features that would be more informative for early prediction. In addition, to further improve the prediction accuracy, future work should also include information on writing profiles to account for individual differences in writing preferences

(C. M. Levy & Ransdell, 1996; Van Waes & Schellens, 2003). Moreover, future work could focus on creating higher-level features, for example, by combining the keystroke data with information derived from natural language processing, to aggregate on the sentence or word level (Leijten & Van Waes, 2013).

Secondly, we included keystroke features which were previously found to be important for the prediction of writing quality, but demonstrated in tasks highly different from the current task. Hence, it could be that these keystroke features do not generalize well to the current task. The literature is also unclear on how the relationship between keystroke data and writing quality differs across tasks. For example, Deane (2014) showed that process and product features of one task could be used to predict another writing task, which indicates that some generalization between tasks is possible. However, Guo et al. (2018) showed that the variance explained by keystroke features differed across six tasks. Therefore, future work should investigate how generalizable the used keystroke features are for the prediction of writing quality across tasks and, specifically of interest for the current study, how generalizable the features are in the context of early prediction of writing quality across tasks.

Lastly, the current approach aims to identify students at risk. However, this is only the first step in addressing students at risk (Romero & Ventura, 2019; Sonderlund et al., 2019). The current approach does not provide any information on the best timing, content, or design of the feedback or computer-based writing support. Although this is out of the scope of the current study, future experimental studies could determine the effect of the content and the design of the feedback on the early writing quality predictions. The keystroke data might also be used to generate the content of the writing feedback. Implications for this based on the current findings are discussed below.

4.4.4 Implications for computer-based writing support

Current computer-based writing support systems mostly assess and provide feedback on the writing product, as opposed to the writing process, and only once the writing is finished (Cotos, 2015; Ma et al., 2014; Wang et al., 2013). In this chapter, we determined whether automated feedback can also be provided during the writing process. Even though we included features that were shown to be related to writing quality in previous studies, our model is not accurate enough for the early prediction of writing quality. However, we contend that providing process feedback during text composition may still be useful.

Therefore, based on the current study, we briefly explore three strands for future research that could further inform us about the effects of process feedback on writing development.

First, we argue that providing students with process feedback during the composition process may improve writers' process awareness. So far, it was quite difficult to create a writing context in which students were challenged to reflect on their writing process. Tutors mainly had to rely on subjective and rather unreliable self-reports. However, thanks to recent developments in keystroke analysis, we can present students with both detailed and global perspectives on different aspects of their writing process, e.g., based on the features extracted in the current study (related to their pausing or revision behavior, fluency, or source usage). By challenging students to reflect on these insights and (possibly) compare their results with their peers, we think we will be able to create a more solid basis to feed the student's process awareness and self-assessment. Follow-up intervention studies that focus on these aspects and measure the effect on process changes and awareness are needed to further support this hypothesis.

Second, we feel it is important to note that there is no such a thing as a 'single' writing process that results in high quality for all students at all times. Writers have individual preferences for specific processes or approaches, which can also be influenced by the task (C. M. Levy & Ransdell, 1996; Van Waes & Schellens, 2003). These preferences are also called writing profiles or writing signatures. The information on the writing process as obtained by the keystroke features can be used to visualize these writing profiles, which can be used for feedback purposes. One example visualization is the progress graph (see also Leijten & Van Waes, 2013), which displays the number of characters produced and the document length as a function of time. Based on the current data, we provide the progress graph of two writers (see Figure 4.5). Here, one writer plans extensively at the beginning of the writing process with many copy-paste actions (top). In contrast, the other writer skips planning in the beginning, but revises extensively at a later stage in the writing process, indicated by the changing position in the document (bottom). Although this results in different writing process characteristics, these writers still show similar writing quality. This already illustrates that there might not be a single 'best' writing process. These kinds of visualizations can be used as the basis for possible interventions, by reflecting on students' processes during the writing task or by comparing progress graphs of a good and a poorer student. A small case study already showed that a feedback report including the progress

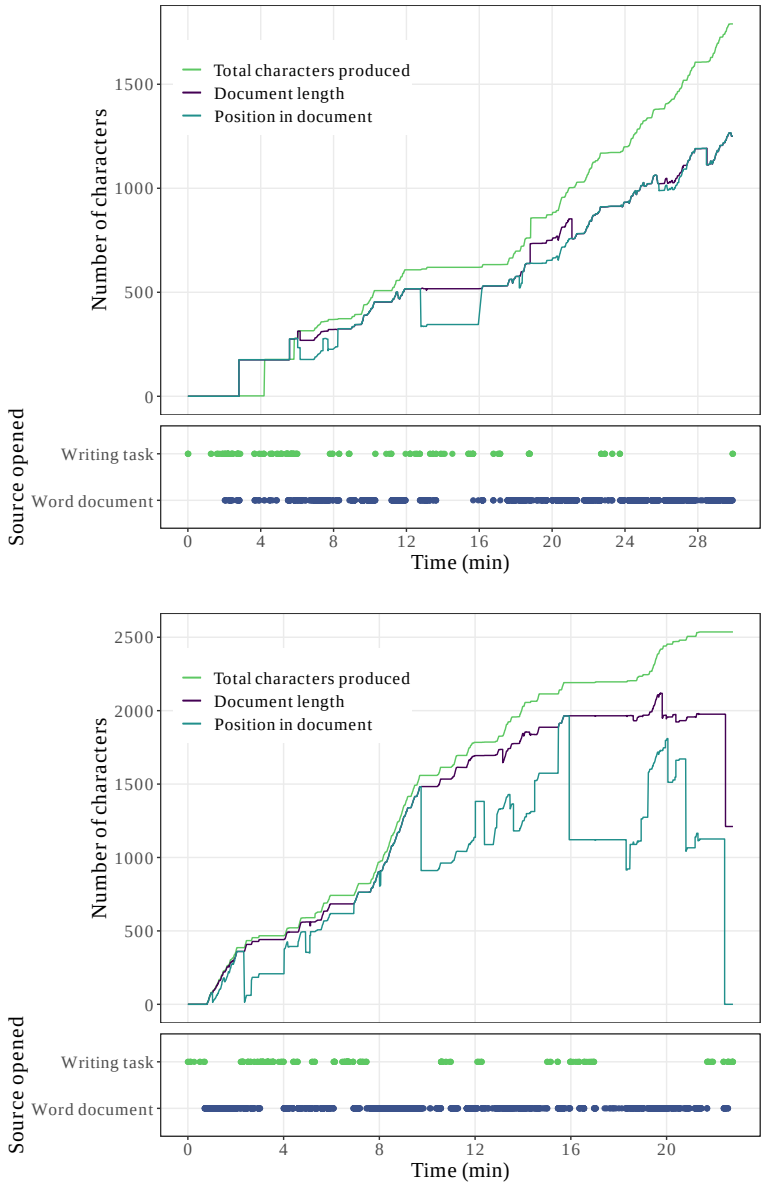

Figure 4.5: Progress graph of two students with similar grades, showing different writing profiles. Top: extensive planning, little revision (participant 10, final grade 9.5); Bottom: little planning, extensive revision (participant 59, final grade 9).

graph helped students to reflect on their writing process and even enabled them to identify ways to improve their writing process (Vandermeulen et al., 2020). Future work should further identify how reflection on the writing process could improve students' writing.

Finally, we showed that the relation between writing quality and the writing process changes over time during the writing process. The correlational analysis showed that features related to other events, such as shifts to task or translation, were correlated with final grade only in the beginning, while features related to fluency, such as percentage of long pauses and percentage of linear transitions between sentences, were only correlated in the middle and end of the writing process. Likewise, the feature importance analysis showed that certain features were only relevant in the middle of the writing process (e.g., percentage of other events or percentage of revision bursts), while others are only important in the beginning (e.g., number of focus shifts to task or mean and *SD* IKI). These findings corroborate with previous studies, which also found that the relation between writing process and quality differs over time, in size and even in the direction of the effect (Breetvelt et al., 1994). This indicates that for providing feedback during writing, we should not only include information keystroke features calculated over the full writing process, but also calculated over different segments within the writing process. In this way, the feedback can be focused on the specific phase the current student is in.

4.5 Conclusion

This chapter provided insight into the relationship between the writing process, measured by keystroke features, and writing quality, measured by final grade. In addition, we determined which of the keystroke features are useful for predicting writing quality during the writing process. In contrast to previous research, the relationship between the writing process and writing quality proved to be rather limited. Hence, these writing process features do not allow for the prediction of writing quality, let alone early prediction. In addition, we showed that the relationship between keystroke features and writing quality changes over time during the writing process. Thus, the relation between keystroke features and writing quality might be less straight-forward than originally posited. However, we contended that providing process feedback during text composition may still be useful and provided pedagogical implications for using keystroke analysis for process-based feedback. Hence, keystroke analysis might be better used to assess the writing process, rather than its product.

5

A product and process oriented tagset for revisions in writing

Adapted from: Conijn, R., Dux Speltz, E., van Zaanen, M., Van Waes, L., & Chukharev-Hudilainen, E. (under review). *A product and process oriented tagset for revisions in writing.*

As we could not find a clear relation between keystroke features and writing *quality*, in the remainder of this dissertation we use keystroke logging to gain insight into the writing *process*, and specifically the revision process. Given the importance of revision in writing, revision has been a main topic of interest in writing research. For providing feedback on revision processes, we first need to collect information on revisions. Several models of revision have been developed, and a variety of taxonomies have been used to measure revision in empirical studies. Current advances in data collection and analysis have made it possible to study revision in more detail. However, a specific approach of how to do this is lacking. Therefore, this chapter aims to provide a comprehensive product-oriented and process-oriented tagset of revisions. The presented tagset consists of ten properties of revisions: processing, trigger, orientation, evaluation, action, linguistic domain, spatial location, temporal location, duration, and sequencing. For each of these properties we detail how features related to these properties can be extracted manually or automatically, using keystroke logging, screen replays, and eye tracking. As a proof of concept, we show how this tagset is used to annotate revisions made by higher education students with various backgrounds in various academic tasks. To conclude, this tagset forms a scalable basis for studying revision in writing in more depth.

Acknowledgements. I would like to thank MacKenzie Novotny, Laura Raught, and Haley Spengler for their assistance in annotating the dataset. The creation of the tagset presented in this chapter was partially funded by Iowa State University's college of Liberal Arts and Sciences (Signature Research Initiative) and the EARLI Emerging Field Group (EarlyWritePro).

5.1 Introduction

Revision has commonly been argued to play an important role in writing (Allal et al., 2004; Flower & Hayes, 1981; Fitzgerald, 1987; Scardamalia & Bereiter, 1983). Revision has been defined as follows: "making any changes at any point in the writing process" (Fitzgerald, 1987, p. 484). Thus, a revision does not necessarily need to correct an *error*; it can be any change within the written text produced so far, as well as a change in the writer's mind before text is written down. Revisions may influence the outcome, such as the writing quality of the written product, the writers' development, such as the writers' knowledge about the topic or about writing, and the writing process (Barkaoui, 2016; Fitzgerald, 1987). Hence, revision is a major area of interest in writing research.

Given the importance of revision in writing research, a variety of models of revision have been developed, sometimes embedded in more general writing process models. One of the earliest writing process models distinguishes *editing* and *reviewing* processes (Hayes & Flower, 1980; Flower & Hayes, 1981). Editing is a change in the text which is automatically triggered, such as a spelling correction, while reviewing is a "systematic examination and improvement of the text" (Hayes & Flower, 1980, p. 18). Reviewing consists of two subprocesses: *evaluating* and *revising*, where evaluating refers to the process of identifying where changes need to be made in the text and revising to the actual change of the text (Flower & Hayes, 1981). Scardamalia & Bereiter (1983) expanded upon this reviewing process with the compare, diagnose, operate (CDO) model (Becker, 2006). In this model, writers *compare* the mental representation of the text written so far with the mental representation of the intended text. If there is a perceived mismatch between these representations, they *diagnose* what needs to be changed and then *operate* this change on the text (Scardamalia & Bereiter, 1983). These three processes have also been identified as the main hurdles in the revision process: detecting the problem, diagnosing the problem, and selecting a strategy (Flower et al., 1986). Flower et al. (1986) represent these hurdles in a more detailed model of revision. This model starts with the task definition, then the writers read their text to comprehend and to evaluate whether their goals are met. This results in a problem representation which can be ill-defined (merely a detection of the problem) or well-defined (a diagnosis of the problem). Based on the problem representation, the problem can either be ignored, or a strategy will be selected: rewrite or revise (Flower et al., 1986).

In addition, several revision taxonomies have been developed to make revisions measurable. These taxonomies tend to be either product-oriented or process-oriented (Lindgren & Sullivan, 2006b). Product-oriented taxonomies focus on the effect of the revision on the writing product, such as the orientation of the revision, e.g., lower-level surface versus a meaning-level (semantic) revisions (Faigley & Witte, 1981). Process-oriented taxonomies focus on the process of making the revision, such as the time and place of the revision in the writing process.

These models and taxonomies have shown to be useful for empirical studies, for example, to determine the differences in revision behavior between skilled and less-skilled writers (e.g., Faigley & Witte, 1981) or the effect of instruction on revision (e.g., Sengupta, 2000). However, these approaches usually only take a few properties of revision into account. In addition, these models, classifications, and taxonomies frequently discuss similar properties but sometimes use different definitions and terminology. Moreover, current advances in data collection and analysis, such as keystroke logging, eye tracking, and natural language processing, have made it possible to gain a more complete and in-depth analysis of revision. Yet, a complete overview of and approach to extracting all these features, especially in relation to the previous models and taxonomies, is lacking.

Therefore, we aim to provide a comprehensive product-oriented and process-oriented tagset of revisions, which can be used for analyzing writing product, such as final texts, and writing process data, such as keystroke logs or text-change logs. For this, we combine the previous taxonomies of revisions, and extend those with information that can be gained from keystroke logging, eye tracking, and natural language processing. As this results in categorical as well as numerical features to describe revisions, we refer to this as a *tagset* of revisions, as opposed to a *taxonomy*. This tagset is measurable, combining manual annotation and automatic extraction of features of revision. In addition, this tagset allows for multiple categories per revision since revisions have properties on different levels, which are not necessarily mutually exclusive (Lindgren & Sullivan, 2006a,b).

As a proof of concept, we show how keystroke logging, screen replays, and eye tracking may be used to manually annotate and automatically extract the features. Here we only extract the features related to revisions in written text (external revisions), as opposed to internal revisions (cf. Murray, 1978), because internal revisions are made before a keystroke is pressed and hence will not be explicitly present in the keystroke data.

5.2 Current revision tagset

In the following, we describe the different properties of revisions in the current revision tagset. These properties are described in relation to the previous models and taxonomies, with a specific focus on how these taxonomies could be used to create a comprehensive tagset that can be used to (manually or automatically) annotate revisions. In addition, we shortly describe how features related to these revisions can be annotated manually or extracted automatically. A more extensive description of how to annotate the manual features can be found in the annotation guide in Appendix B.

5.2.1 Revision events

Before undertaking the task of annotating revisions, we first need to define a *revision event*, i.e., a specific episode of the writing process where a single revision takes place. This then becomes our unit of analysis.

For the purposes of the present tagset, we consider that a new revision event begins when one of the following takes place:

1. The writer begins deleting characters in the text.
2. The writer moves the cursor to a different location in the text and then begins producing new characters.

The beginning of the revision event is identified automatically based on the analysis of the writing process data (i.e., keystroke logs and text-change logs). The revision event is considered finished when one of the following takes place:

1. The writer initiates a new revision event as defined above.
2. Upon finishing the revision, the writer continues with text production.

In the first case, the end of the revision event is marked automatically. However, if a revision event is followed by an episode of new text production (second case), which is not directly caused by the revision, the end of such revision event is marked manually by dataset annotators.

5.2.2 Revision properties

After identifying the revision event, each revision event is described using a set of properties. In total, we identified ten categories of properties addressed in the literature for the current tagset: processing, trigger, orientation, evaluation, action, linguistic domain, spatial location, temporal location, duration, and sequencing. For each of the properties, we

identified a wide variety of features related to these properties. For the current tagset, we specifically restricted the features to be *measurable* features, that is, features that can be automatically extracted or features that can be annotated manually, using writing process data. In addition, we choose features that are not related to a specific genre. An overview of all features identified related to each of these properties can be found in Table 5.1.

Table 5.1: Features in the revision tagset per property

Property	No.	Feature	Type
A. Processing	1	Processing (internal pre-linguistic, internal pre-textual, external)*	categoric
B. Trigger	1	Feedback type (no feedback, spelling, grammar, punctuation, other)*	categoric
C. Orientation	1	Surface change	binary
	1.1	Typography	binary
	1.2	Capitalization	binary
	1.3	Punctuation	binary
	1.4	Spelling	binary
	1.5	Grammar	binary
	1.6	Cosmetics/presentation	binary
	1.7	No change	binary
	1.8	Wording/phrasing	binary
	2	Semantic change	binary
	2.1	Microstructure change (supporting info, emphasis, understate, coherence, cohesiveness)	categoric
	2.2	Macrostructure change (overall aim, subtopic)	categoric
D. Evaluation	1	Correct start	binary
	2	Correct finish	binary
E. Action	1	Action (insertion, deletion, substitution, reordering)	categoric
F. Domain	1	Linguistic domain (subword, word, phrase, clause, sentence, paragraph)	categoric
	2	Number of backspaces	numeric
	3.1	Number of characters deleted	numeric
	3.2	Number of characters inserted	numeric
	4.1	Number of words deleted	numeric
	4.2	Number of words inserted	numeric

Table 5.1: Features in the revision tagset per property (continued)

Property	No.	Feature	Type
	5.1	List of POS tags words deleted	list
	5.2	List of POS tags words inserted	list
	6.1	List of chunk tags words deleted	list
	6.2	List of chunk tags words inserted	list
	7.1	Number of sentences deleted	numeric
	7.2	Number of sentences inserted	numeric
G. Spatial location	1	Word finished	binary
	2	Intended word	string
	3.1	Word initial	binary
	3.2	Clause initial	binary
	3.3	Sentence initial	binary
	4	Number of characters from leading edge	numeric
	5	Number of words from leading edge	numeric
	6	Pre-contextual/contextual	binary
	7	Immediate/distant	binary
	8.1	Number of characters from start sentence	numeric
	8.2	Number of chars. from start writing process	numeric
	8.3	Number of chars. from start writing product	numeric
H. Temporal location	1	Time from start writing process	numeric
I. Duration	1	Duration of the revision event	numeric
	2	Pause time before revision	numeric
J. Sequencing	1	Overrides previous	binary
	2	Continues previous	binary
	3.1	Repetitive (leading edge)	binary
	3.2	Repetitive (immediate)	binary
	4.1	Embedded (leading edge)	binary
	4.2	Embedded (immediate)	binary
	5.1	Sequence forwards (leading edge)	binary
	5.2	Sequence forwards (immediate)	binary
	6.1	Sequence backwards (leading edge)	binary
	6.2	Sequence backwards (immediate)	binary
	7	Time from previous revision	numeric
	8	Number of characters from previous revision	numeric

Note. * This feature was not extracted in the current chapter.

Here, the properties orientation, evaluation, action, and domain could be considered product-oriented, as these focus on the effect of the revision on the writing product, whereas processing, trigger, spatial location, temporal location, duration, and sequencing are process-oriented, as these focus on the process of making the revision. Below we describe the theoretical background and implementation of each of the properties in detail.

A. Processing. Revisions have been divided into two subprocesses: reading and editing Hayes & Flower (1980). The editing process can be further divided into four modes: editing to adhere to writing conventions, editing to improve semantics, evaluating to improve readers' understanding, and evaluating to improve readers' acceptance (Hayes & Flower, 1980). One year later, Flower & Hayes (1981) redefined the subprocesses of reviewing into evaluating and revising. Here, evaluating can include different modes, such as evaluating the text written so far as well as evaluating the writers' planning or their unwritten thoughts and statements (Flower & Hayes, 1981). Other researchers explicitly distinguished the process of revision in these two modes: internal and external revisions, mostly measured using thinking-aloud. Internal revisions are defined as mental revisions that are made before transcription, while external revisions are revisions that are visible in the written production (Murray, 1978). In later studies, internal revisions were further subdivided into pre-linguistic revisions, or changes made to non-linguistic mental representations, and pre-textual revisions, or revisions made to linguistic mental representations, which could both affect conceptual content as well as formulation (Lindgren & Sullivan, 2006b; Stevenson et al., 2006).

Accordingly, for the property processing, we included a categorical feature consisting of pre-linguistic and pre-textual internal revisions, and external revisions. In the current chapter, processing is not annotated, as the dataset did not include think-aloud protocols that would allow us to annotate internal revisions.

B. Trigger. The trigger of the revision describes the cause of the revision. The triggers identified in the literature include reading and evaluating the text written so far, or reading the writing task and evaluating the written text (Tillema et al., 2011). Other researchers focus on revisions triggered by errors (Stevenson et al., 2006), also known as conventional revisions—revisions that are necessary to correct to fulfill linguistic requirements (Allal, 2000)—in contrast to optional or non-error-triggered revisions. With the advent of automatic writing evaluation systems, a revision can also be triggered externally, i.e., not by the

writer itself. One common example is spelling or grammar checkers which trigger revisions in spelling or grammar (Figueredo & Varnhagen, 2006). The trigger of the revision is usually highly related to the orientation or the depth of the revision (see below). Therefore, the revision trigger is sometimes categorized as revision orientation (e.g., revision triggered by a typo versus typo revision).

Accordingly, we categorized trigger mostly under orientation, except for revisions triggered by feedback. For trigger, one automatic feature is included: type of feedback (no feedback, spelling, grammar, punctuation, and other). In the current chapter, trigger is not extracted, as no feedback was provided in the current dataset.

C. Orientation. One of the most frequently used classifications of revision involves the orientation or the depth of the revision (Faigley & Witte, 1981). Faigley & Witte's (1981) taxonomy on revisions distinguishes surface and semantic revisions. Surface revisions are further subdivided into formal changes (spelling: tense, number, and modality; abbreviations; punctuation; and format) and meaning-preserving revisions that paraphrase, but do not alter the meaning of concepts in the text. Semantic changes have a more fundamental impact on the actual text and can be divided into minor (microstructure) and major (macrostructure) changes which alter the summary of the text. Meaning-preserving, microstructure, and macrostructure revisions are all further divided into additions, deletions, substitutions, permutations, distributions, and consolidations. In the current tagset, we treat these subcategories separately under revision action.

Other studies used variations of Faigley & Witte's (1981) taxonomy on revision changes. For example, in Lindgren & Sullivan's (2006a) taxonomy, they classified contextual revisions, revisions away from the leading edge of the text produced so far (see spatial location), into form and conceptual revisions. Form revisions are similar to Faigley & Witte's (1981) surface revisions but are further extended into typography (slip of the finger, see also Chapter 6), spelling, grammar, punctuation, formatting, and meaning-preserving revisions. Several subcategories were further specified: spelling (revision, substitution, deletion, and homophone), grammar (e.g., verb tense, subject/verb agreement, preposition, conjunction, article, pronoun, genitive, adverb, and other), and format (punctuation, capitalization, paragraph, and other format). In addition, changes in language (L2 to L1 and L1 to L2) were added as meaning-preserving changes. The conceptual revisions are based on Faigley & Witte's (1981) semantic revisions and consist of content-based revisions (microstructure

and macrostructure changes) and balance revisions, related to the topic or audience (register, other). Microstructure changes are changes that affect meaning but would not alter the gist of a text, whereas macrostructure changes are changes that would alter the summary of a text. Allal (2000) chose a slightly different approach and distinguished between semantics (lexical variations and changes of meaning), text organization (segmentation, connection, and cohesion), and spelling (lexical and grammatical aspects). Here, the first two could be considered semantic changes, while the latter constitutes surface changes. Sometimes, the orientation labels for the semantic changes are tailored to a specific genre; e.g., for argumentative writing, F. Zhang et al. (2016) distinguished semantic changes in thesis/idea (claim), rebuttal, reasoning (warrant), evidence, and general content changes.

A few studies report on functions or purposes of revisions which we argue would be closely related to the orientation of the revision. For example, Monahan (1984) identified five purposes: cosmetic, mechanical, transitional, informational, and stylistic, where cosmetic, mechanical, and stylistic could be considered as surface changes, and transitional and informational as semantic changes. Similarly, Falvey (as cited in Min, 2006) distinguished grammatical, cosmetic (surface changes), texture (increasing cohesion and coherence), unnecessary expression, and explicature (increasing explicitness) revisions.

For the current tagset, primarily, we distinguish between surface and semantic (or deep) revisions. Surface revisions include conventional copy-editing operations or revisions that paraphrases a concept in the text, but do not alter it. Surface changes are divided into formal changes (typography, spelling, grammar, capitalization, punctuation, cosmetics/presentation, and no change) and meaning-preserving changes (wording/phrasing). All these categories were manually annotated, except for punctuation, capitalization, and no change, because these could be automatically extracted based on the type of the character inserted or deleted. Typography and spelling were distinguished by applying the guidelines used in previous manual annotations (Stevenson et al., 2006; Wengelin, 2007). If it was unclear which subcategory should be selected, all possible orientations needed to be selected (e.g., both wording and typographic error when only the first letter of a word was revised). Semantic or deep revisions, as opposed to surface revisions, do change the semantics or meaning of the text. We divided semantic changes into microstructure and macrostructure changes. For microstructure changes, we further distinguished adding or removing supporting information, changing emphasis, understatement, coherence, and

cohesiveness. Macrostructure changes consisted of altering the overall aim and adding or removing entire subtopics.

D. Evaluation. The evaluation of the result of a revision has typically been used for formal changes only: surface changes necessary to fulfill linguistic requirements (e.g., spelling, grammar, punctuation). Possible evaluations that have been used include correct or incorrect (Allal, 2000); correct, incorrect, or neutral (Chanquoy, 1997); or successful or unsuccessful revision (Stevenson et al., 2006). Moreover, some studies described whether the starting point was correct: in other words, whether an error or non-error was revised (e.g., from a correct non-error into an error; Wobbrock & Myers, 2006). By contrast, Crawford et al. (2008) evaluated the quality of all revisions. As it is harder to determine whether semantic changes are "correct," they renamed the labels into increase, decrease, and neutral.

In the current tagset, the evaluation of the revision was manually annotated. This was only done when the revision was oriented toward typography, spelling, or grammar, as these types of formal changes are necessary for satisfying linguistic requirements, and hence can be evaluated as correct or incorrect. Other changes, such as semantic changes, cannot be easily assessed, as these would require information on the context surrounding the revision, as well as information on the requirements of the writing task. In the current tagset, we both describe whether the starting point was correct or incorrect and whether the result of the revision was correct or incorrect.

E. Action. Most of the actions or revision operations stem from work on the classification of single letter errors which consists of an extra letter (insertion), a missing letter (omission), a wrong letter (substitution), and a transposition of two adjacent letters (transposition; Damerau, 1964). These have been translated into four revision actions: deletion, insertion, substitution, and reordering (also referred to as transposition, reorganization, and permutations), respectively (e.g., Allal, 2000; Sommers, 1980). These actions could involve a single letter but also a larger linguistic unit (linguistic domain, see below). Some researchers added one or more actions to these four actions. For example, Monahan (1984) added embedding, a specific form of insertion, where new information is embedded within the structure of already written text. Faigley & Witte (1981) included distribution, where a single unit is changed into more than one unit (e.g., one sentence is distributed into two), and its opposite, consolidation, where two or more units are combined into one unit.

Here, the revision action was automatically classified into insertion, deletion, substitu-

tion, and reordering (transposition). For the revisions below word level, the revision action was determined by the edit distance (restricted Damerau-Levenshtein distance; Boytsov, 2011) of the deleted and inserted text (up to the manually annotated revision end). For the revisions at word level and above, the revision was coded as insertion if more words were inserted than deleted; as a deletion if more words were deleted than inserted; as a transposition if the same words were replaced, but in a different order; and as a substitution otherwise.

F. Linguistic domain. Several studies included the linguistic domain affected by the revision, also known as the level, scope, range, or size of the revision. Monahan (1984) distinguished subword (surface), word, phrase, clause, sentence, paragraph, and discourse level, while Stevenson et al. (2006) only distinguished below word, below clause, and clause and above. Some researchers focus more on smaller domains, e.g., sub-grapheme, grapheme, morpheme, word, and above-word (Lindgren et al., 2019). Others added domains such as theme (extended statement of one idea; Sommers, 1980), punctuation (Crawford et al., 2008), or revisions made within the current functional component being realized (forward progressions) versus revisions made within the current component, but across component boundaries (communicative progression; Bowen & Van Waes, 2020). Xu (2018) took a lower-level approach by merely counting the characters deleted or produced.

In the current tagset, the linguistic domain was manually classified into one of the following six categories: revision within a subword, word, phrase, clause, sentence, or paragraph. In addition to the manual features, several automatic features were extracted: the number of backspaces, characters, words, and sentences inserted and deleted. The number of characters, words, and sentences inserted was calculated up to the manually annotated revision end. Note, the number of backspaces and the number of characters deleted are related, but will differ, for example when different revision techniques are used (e.g., selection of ten characters and one backspace keystroke versus ten backspace keystrokes). Finally, if one or more words were deleted or inserted, we also extracted a list of part-of-speech tags of the words deleted or inserted, using the part-of-speech tagger in the Natural Language Toolkit in Python (Bird et al., 2009). Likewise, when more than one word was deleted or inserted, we extracted a list of chunk tags of the words deleted or inserted, using the chunk tagger in the Natural Language Toolkit in Python (Bird et al., 2009).

G. Spatial location. Revisions occur at different locations in the text produced so

far. We refer to this as the spatial location in the writing process. We distinguished two slightly different operationalizations of the spatial location in the writing process. On the one hand, Lindgren & Sullivan (2006a,b) distinguished pre-contextual and contextual revisions. Pre-contextual revisions are defined as revisions at the point of inscription or leading edge of the text produced so far. Contextual revisions are defined as revisions made when the writer moves away from the leading edge and makes a revision in a previously written and completed sentence. In this way, revisions away from the leading edge but in an uncompleted sentence are left uncategorized. Accordingly, Baaijen et al. (2012) used a simpler categorization and only distinguished between revisions at the leading edge and revisions elsewhere in the text. On the other hand, Thorson (2000) distinguished immediate (or intermediate) and distant revisions. Here, immediate revisions are revisions at the point of the cursor location, and distant revisions include revisions where the cursor moved away to start a revision elsewhere in the text. For distant revisions, the distance from the cursor location has been included, such as the number of lines below or above the cursor location (Van Waes & Schellens, 2003). Likewise, a distinction has been made between movements from the cursor location forward and backward in the text (Bowen & Van Waes, 2020). Xu (2018) added a third category to immediate and distant revisions: end revisions, which are revisions made after the completion of the whole text. In the current tagset, we treat this third category separately under temporal location in the writing process.

Several features related to the spatial location were automatically identified. First, we distinguished pre-contextual and contextual revisions. As Lindgren et al. (2019) recommended, we also included revisions before the last character(s) in the written text if the last character(s) only consisted of invisible characters (e.g., a trailing white space). However, sometimes a writer first starts with a rough outline (e.g., intro-body-conclusion) and then starts to fill in the gaps. Then, there will almost always be characters behind the cursor, and all revisions will be contextual. Therefore, we also extracted immediate versus distant revisions. In addition, we included the number of words and characters from the leading edge (cf. Van Waes & Schellens, 2003). Lastly, the number of words from the start of the sentence, the start of the writing product (the number of words in the written product up to the revision), and the start of the writing process (the number of words typed so far) were included.

In addition, we added several manual features of spatial location not previously dis-

cussed in revision taxonomy literature. Often, corrections are made when a word is not yet finished. This makes it hard to manually annotate the revision, as it might be impossible to know what the writer was supposed to write, especially if only one or a few characters were typed (Lindgren et al., 2019). This is also hard for computers, as half-written words are especially hard to parse, for example, for part-of-speech tagging. Therefore, as a metric of uncertainty of the manual annotations, the annotators indicated whether the word in which the revision started was finished. If the word was not finished, the intended word was guessed (if possible). Lastly, the location up to where the characters are deleted, or where characters are inserted, could provide information on the orientation of the revision. For example, when characters are deleted up to the middle of a word, it would be less likely to be a semantic change compared to when characters are deleted up to the start of a sentence. Therefore, we also included whether the location of the revision was word initial, clause initial, and/or sentence initial.

H. Temporal location. Temporal location refers to the point in time when the revision is made. Temporal location has been defined both within and between drafts. Monahan (1984) described the temporal locations between drafts: pre-writing stage, during the first draft, between drafts, during the second draft, and after the second draft. Others described the temporal location within a draft. Tillema et al. (2011) split the writing process into episodes of equal durations. With three episodes, for example, you could identify whether the revision took place in the begin, middle, or end of the writing process (Barkaoui, 2016; Van Waes & Schellens, 2003). Others only distinguished revisions made during text production and revisions made at the final writing stage after completing the whole text (Xu, 2018). Lastly, some researchers used a more computational approach and defined the temporal location as the time elapsed from the start of the writing process (M. Zhang et al., 2016).

For the temporal location of the revision, only one automated feature was extracted: the time (in milliseconds) until the revision from the start of the writing process. Combined with the total time, the time until revision can be used to determine any episode in which the revision took place (e.g., begin, middle, end). In addition, combined with the version number of the document, this can be used to determine any time within a writing session (e.g., end of first draft, begin of final version).

I. Duration. One (rather technical) aspect of revision that has been included is the

duration of the revision. The duration is usually measured in milliseconds or seconds from the start of the revision until the end of the revision (Xu, 2018). Duration has also been expressed as the length of an R-burst, where an R-burst is a sequence of text production involving a revision and bounded by long pauses (e.g., pauses > 2 seconds; Chenoweth & Hayes, 2001).

For the duration of the revision, two features were automatically extracted. First, the duration of the revision was calculated by extracting the time from the first editing keystroke until the last editing keystroke. The last editing keystroke was extracted from the manually annotated revision end. Second, the pause time before the revision was extracted: the time from the last key press before a revision until the first key press of the revision.

J. Sequencing. All previous properties of revision are related to a single revision event. Sequencing describes the relation of a revision to a preceding revision in time. As Lindgren & Sullivan (2006a) describe, some revisions might be single, independent revisions, while other revisions might be part of a series of revisions, also known as revision episodes. As examples, they describe three types of revision episodes, as suggested by Kollberg (1996). First, they discuss episodes of revision at one cursor location. For example, a writer may start a sentence, delete it, start it again, delete it again, and then write the final full sentence. In this case, the start of the sentence (same cursor location) is revised twice. Second, they describe episodes with embedded revisions, where a revision is made within a previous revision. Third, they note episodes with a sequence of revisions. For example, to maintain consistency, one might change the wording or spelling of a specific word and change that word throughout the text. Williamson & Pence (1989) also discussed a fourth category, where a revision "inspires" another revision without reviewing the text. For example, a revision of the last sentence of a paragraph might be followed by a revision in the beginning of that paragraph. Another approach to examining the sequencing of revisions is the S-notation, which maps all deletions and insertions in the text to the spatial location in the writing process (Kollberg, 1996). With this S-notation, the non-linearity of the writing process can be analyzed, and connected episodes of revision can be automatically identified (Kollberg, 1996; Severinson-Eklundh & Kollberg, 2001).

In the current tagset, four types of sequencing were automatically identified: repetitive revisions, embedded revisions, sequence forward revisions, and sequence backward revisions. Repetitive revisions are operationalized as revisions which start (in the writing

product) within two characters from the start (in the writing product) of the previous revision. Embedded revisions are revisions where the range of character positions of the start and end of the revision (within the writing product) fall within the range of character positions of the previous revision. Sequence forward includes revisions where the subsequent revision was made further on in the text (forward from the last revision), while sequence backward includes revisions where the subsequent revision was made earlier in the text. All four types were calculated based on the leading edge (pre-contextual/contextual) and cursor (immediate/distant) spatial locations. Lastly, we automatically extracted metrics related to the distance from the previous revisions: time and number of characters from the previous revision.

In addition to the automatic features, we added two manual features: overrides previous revision and continues on previous revision. A revision overrides the previous revision if it is repetitive, i.e., changing the same linguistic domain at the cursor location. A common example of this is when a writer makes a typographical error and attempts to revise it, but then makes a typographic error again in that revision attempt. A revision continues on a previous revision if the previous revision caused the subsequent revision. For example, changing a word from singular to plural might result in changing the verb to maintain subject-verb agreement.

5.3 Proof of concept

As a proof of concept, we used our tagset to describe the revisions made by university students while conducting an academic writing task. The manual features were manually annotated by multiple raters, and the other features were automatically extracted. The features related to trigger were excluded because no feedback was provided in the current dataset; hence the revisions could not be triggered by feedback.

5.3.1 Dataset

Our dataset was created by sampling writing-process data from a large data store containing anonymized (i.e., not containing any personally identifiable information or links to research participants) writing-process log files. These log files were recorded as part of various prior research studies conducted using CyWrite, a web-based word-processing tool that collects keystroke data and eye fixations during writing (Chukharev-Hudilainen et al., 2019; Chukharev-Hudilainen, 2019; Ranalli et al., 2018a). In addition, the CyWrite tool

offers playback functionality to replay the screen recording combined with the gaze-point marker. To include a wide variety of revisions, a stratified random sample was collected from the data store, which ensured a diversity of tasks and writer backgrounds.

This resulted in data from 20 native English graduate students writing four different tasks (all of which prompted the students to write a 150–250 word abstract of a research article); 20 native English undergraduate students writing two different tasks (both of which were argumentative tasks adapted from the Test of English as a Foreign Language); and 25 English Second Language learners (most likely undergraduate students based on the original study that contributed to this portion of the dataset) writing the same two tasks as the native English undergraduates.

These 65 participants had a total of 7,120 revision events (M = 110, SD = 53). For every revision event, all manual and automatic features listed in Table 5.1 were extracted, except for the processing and trigger features.

5.3.2 Feature extraction

For the manual annotations, a spreadsheet was created with (for every revision event; i.e., for every row) the revision event id, removed characters, and typed characters (columns). Next to this information, the annotator could further explore the revision event (in the context of the writing process) using the visual playback function in the CyWrite interface, providing a high-fidelity, keystroke-by-keystroke animated reconstruction of the text production process with an overlaid eye fixation marker. For all features that needed to be manually annotated, we created an extensive annotation guide. Within this guide, guidelines, explanations, and examples are provided for each label. This annotation guide was first created by the authors through several rounds of discussion, trial coding, and evaluation. Thereafter, four annotators were trained to manually annotate using the annotation guide. In each training round, we explained the guidelines to the annotator, the annotator coded a sample document, disagreements were discussed, and where necessary, we clarified the annotation guide with additional explanations or examples. The final annotation guide including annotation examples can be found in Appendix B.

Table 5.2: Descriptive statistics of all features and inter-rater reliability (Krippendorff's alpha) of the manual annotated features in the revision tagset, *N* = 7,120

Property		Feature	Mean (*SD*)	IRR (α)
General	1	Revision [Y/N]	91.9% (3.7%)	0.96
	2	Revision end[a]		0.74
C. Orientation	1	Surface	92.6% (6.2%)	0.64
	1.1	Typography	50.8% (13.0%)	0.71
	1.2	Capitalization	1.6% (2.2%)	.
	1.3	Punctuation	6.4% (3.5%)	.
	1.4	Spelling	2.6% (3.1%)	0.74
	1.5	Grammar	9.0% (4.9%)	0.69
	1.6	Cosmetics/presentation	0.2% (0.6%)	0.83
	1.7	No change	7.4% (3.8%)	.
	1.8	Wording/phrasing	21.0% (10.9%)	0.75
	2	Semantic (deep)	13.9% (8.6%)	0.59
	2	Deep specify[b]		0.22
	2.1	Microstructure changes	14.1% (8.6%)	.
	2.1.1	Supporting info	6.9% (4.8%)	.
	2.1.2	Emphasis	2.0% (2.3%)	.
	2.1.3	Understate	0.8% (1.1%)	.
	2.1.4	Coherence	1.4% (2.0%)	.
	2.1.5	Cohesiveness	0.4% (0.8%)	.
	2.1.6	unknown	2.6% (2.9%)	.
	2.2	Macrostructure changes	0.0% (0.2%)	.
	2.2.1	Overall aim	0.0% (0.0%)	.
	2.2.2	Subtopic	0.0% (0.2%)	.
D. Evaluation	1	Correct start	4.7% (4.2%)	0.69
	2	Correct revision	85.2% (9.4%)	0.66
E. Action	1.1	Insertion	40.0% (15.5%)	.
	1.2	Deletion	25.2% (7.6%)	.
	1.3	Substitution	24.4% (10.3%)	.
	1.4	Reordering	3.0% (2.7%)	.
F. Domain	1	Domain specify[b]		0.59
	1.1	Subword	67.7% (11.6%)	.
	1.2	Word	24.1% (9.2%)	.
	1.3	Phrase	4.6% (4.0%)	.
	1.4	Clause	1.3% (1.5%)	.
	1.5	Sentence	2.2% (3.1%)	.
	1.6	Paragraph	0.0% (0.2%)	.

Table 5.2: Descriptive statistics of all features and inter-rater reliability (Krippendorff's alpha) of the manual annotated features in the revision tagset, *N* = 7,120 (continued)

Property		Feature	Mean (*SD*)	IRR (α)
	2	Number of backspaces	2.4 (0.5)	.
	3.1	Number of characters deleted	3.5 (1.4)	.
	3.2	Number of characters inserted	8.2 (8.7)	.
	4.1	Number of words deleted	1.1 (0.2)	.
	4.2	Number of words inserted	1.7 (1.5)	.
	7.1	Number of sentences deleted	0.02 (0.02)	.
	7.2	Number of sentences inserted	0.04 (0.09)	.
G. Spatial location	1	Word finished	51.0% (12.0%)	0.70
	2	Intended word[a]		0.71
	3.1	Word initial	43.1% (10.9%)	0.80
	3.2	Clause initial	13.7% (6.7%)	0.68
	3.3	Sentence initial	10.2% (6.3%)	0.82
	4	Characters from leading edge	69.3 (91.4)	.
	5	Words from leading edge	11.6 (15.7)	.
	6	Pre-contextual (= 1 - contextual)	77.9% (17.1%)	.
	7	Immediate (= 1 - distant)	86.2% (10.1%)	.
	8.1	Characters from start sentence	69.2 (31.7)	.
	8.2	Chars. from start process	814 (349)	.
	8.3	Chars. from start product	817 (345)	.
H. Temporal location	1	Time from start process (min)	8.5 (4.0)	.
I. Duration	1	Duration (sec)	3.1 (3.0)	.
	2	Pause before revision (sec)	2.0 (1.1)	.
J. Sequencing	1	Overrides previous revision	13.8% (7.6%)	0.55
	2	Continues on previous revision	14.6% (8.3%)	0.27
	3.1	Repetitive (leading edge)	23.9% (10.1%)	.
	3.2	Repetitive (immediate)	23.8% (10.4%)	.
	4.1	Embedded revision (leading edge)	0.2% (0.5%)	.
	4.2	Embedded revision (immediate)	0.2% (0.5%)	.
	5.1	Sequence forwards (leading edge)	8.1% (8.3%)	.
	5.2	Sequence forwards (immediate)	4.8% (5.2%)	.
	6.1	Sequence backwards (leading edge)	1.4% (2.1%)	.
	6.2	Sequence backwards (immediate)	1.2% (1.8%)	.
	7	Time from previous revision (sec)	6.7 (3.4)	.
	8	Characters from previous revision	7.0 (11.1)	.

Note. [a] Non-numerical variable so no descriptive statistics can be provided. [b] Inter-rater reliability is calculated once for the full category, as all labels are mutually exclusive. A "."-symbol indicates a feature that is automatically extracted, so no inter-rater reliability available.

After each training round, the inter-rater reliability was used to determine whether more training was needed. In addition, this showed which features specifically required the most attention, that is, which features received the lowest reliability. Krippendorff's alpha (α; Krippendorff, 2011) was used as inter-rater reliability metric, as it allows for multiple raters, multiple measurement levels (e.g., numerical and categorical) and is more reliable for coding sparse categories, as it focuses on the disagreement rather than the agreement. We considered an annotator fully trained when the inter-rater reliability of all features between author and annotator were similar to the inter-rater reliability between two authors. In total, two to three rounds were needed to train each annotator. After the training rounds, the annotators independently annotated the documents. Of the 65 documents annotated, 15 (23%) documents were randomly selected to be annotated twice (by different pairs of annotators) to calculate the inter-rater reliability of the fully annotated dataset. The inter-rater reliability of the manual annotations (after training) can be found in Table 5.2. The automatic features were extracted using JavaScript and R as described in Section 5.2. The annotated dataset can be found at Conijn, Dux Speltz, et al. (2020).

5.3.3 Results

The descriptive statistics of the manually annotated and automatically extracted features are shown in Table 5.2. Regarding the orientation of the revision, most revisions were surface revisions (92%), and especially typographic revisions or wording/phrasing revisions. Only 14% of the revisions were deep revisions. Note that the total is more than 100% here, which indicates that several revisions are annotated both as surface and as deep revision. Thus, some for some revisions, it was not visible in the keystroke data and screen replays with eye fixation whether the revision was a surface or deep revision. For the spelling, grammar, punctuation, and capitalization revisions, most were correctly revised (85%). For some of the revisions, the previous form was already correct (5%). Most revisions were insertions (40%), one-fourth of the revisions were deletions, and one-fourth were substitutions. Only 3% of the revisions involved reordering of characters or words. The domain of the revisions was rather small. Two-thirds of the revisions involved only a few characters, and 24% of the revisions involved a word. Only 8% of the revisions were larger than a word. This was also shown in the low number of characters inserted and deleted, on average. Half of the time, the revisions were made while the current word typed was not finished yet. In

43% of the revisions, the revision deleted up to a word initial position or inserted at the start of a word.

Most revisions were made at the leading edge (pre-contextual revision) and at the point of inscription (immediate). The revisions were spread out over the full writing process, started after a pause of 2 seconds, and took 3 seconds, on average. A substantial amount of the revisions was found to follow-up on previous revisions, with 24% of the revisions repetitive, 8% sequencing forward based on the leading edge, and 5% sequencing forward based on the cursor position. Only 1% of the revisions was sequencing backwards, and less than 1% consisted of embedded revisions.

The inter-rater reliabilities of the manual annotations (see Table 5.2) show that, regardless the training of the annotators, some features were still hard to annotate. For orientation, the overarching category ($\alpha = 0.64$) and the subcategories ($\alpha = 0.71–0.83$) of surface revision proved to be easier compared to the overarching ($\alpha = 0.59$) and subcategories ($\alpha = 0.22$) of semantic revisions. Furthermore, spatial location proved to be easiest to annotate, followed by evaluation and domain. Sequencing proved to be really hard to annotate, and especially the feature continues on previous revision ($\alpha = 0.27$).

5.4 Discussion

In this chapter, we aimed to provide a comprehensive product and process oriented tagset of revisions. The tagset includes ten properties to describe revisions: processing, trigger, orientation, evaluation, action, linguistic domain, spatial location, temporal location, duration, and sequencing. As a proof of concept, we showed how keystroke logging, screen replays, and eye tracking could be used to both manually annotate as well as automatically extract features related to these properties. A few of those properties are discussed in more detail below.

Orientation is one of the most frequently discussed properties in the literature; however, it is currently mostly annotated manually, which is a time-intensive task. The relatively low inter-rater reliability, especially for the deep revisions, also showed that regardless of the training, orientation is hard to annotate. Previous studies showed somewhat higher levels of inter-rater reliability (e.g., Barkaoui, 2016; Stevenson et al., 2006); however, these used different metrics and hence are hard to compare. One study also reported Krippendorff's alpha and showed similar reliability for the surface revisions (Daxenberger & Gurevych, 2012). The lower reliability for the deep revisions can be explained by the

fact that deep revisions were rather sparse in our dataset. Krippendorff's alpha penalizes sparse categories more, and in addition, few deep revisions were available in datasets for training the annotators. Moreover, the lower reliability could indicate that the nature of a deep revision might be too hard to identify with screen replays and eye tracking alone. Triangulating the data with direct input from participants (e.g., through concurrent think-aloud or stimulated recall) might be necessary to get a better understanding of the writers' intentions behind these deep revisions.

For spatial location, two approaches were used: pre-contextual versus contextual revisions (Lindgren & Sullivan, 2006a,b) and immediate versus distant revisions (Thorson, 2000). Although the approaches are quite similar, in 9% of the cases an immediate revision was a contextual revision, and in 7% of the cases a distant revision was a pre-contextual revision. This indicates that these approaches are still quite different and cannot be used interchangeably. We would argue that the 'best' approach depends on the research question and the writing task at hand. For example, in tasks where students are required to make an outline up front, the immediate versus distant approach might work better, as otherwise many revisions would be considered contextual once students start to fill out the gaps in the outline. In addition, we included the number of characters away from the leading edge. This allows for a more detailed approach: the leading edge might be updated when the number of characters away from the leading edge stay the same for several revision events. In this way, not only invisible characters (cf. Lindgren et al., 2019) are ignored for leading edge calculations, but also visible but perhaps currently 'unused' characters, such as a trailing dot or the bibliography.

The sequencing of the revision was analyzed using both manual and automated features. However, the inter-rater reliability of the manual annotations was low. The annotators indicated that it was sometimes hard to determine whether a revision continued a previous revision, especially when it was interrupted by one small (e.g., typo) revision. Also, the intended word was not always known, making it hard to identify whether the revision followed up on a previous revision. Luckily, a wide variety of automated features are available, potentially making the manual annotation of sequencing unnecessary in many cases. However, these automated features also only determine the sequencing in terms of the immediately preceding or following revision, rather than over multiple revisions. More complex analysis involving natural language processing and S-notation representa-

tions (Kollberg, 1996; Severinson-Eklundh & Kollberg, 2001) could be used to further investigate connected revision episodes over multiple revisions (Leijten, Van Horenbeeck, & Van Waes, 2019). Alternatively, features from the other properties could be used to identify sequences of revisions. For example, the characters from the start of the writing product (spatial location) could be used to determine whether revisions are made within the same sentence or word.

In general, a reasonable inter-rater reliability was reached for most of the manual features. The eye-tracking data added to replays of writing processes proved to be especially useful for making annotation decisions about some specific categories. For example, an eye fixation on a similar word previously written was a clear indication of a spelling revision (as opposed to a typographic revision). Likewise, an eye fixation on a previously written word could provide information on the intended word. For certain categories eye fixation was not really necessary (e.g., evaluation, action, temporal location, and duration), indicating that, depending on which categories are needed, eye fixation does not always need to be extracted.

5.4.1 Limitations

Although we did aim for a comprehensive tagset, the tagset does not consist of context-specific features. For example, F. Zhang & Litman (2015) included tags related to the genre, such as changes in warrant, reasoning, or backing for argumentative writing. Therefore, the tagset cannot be used to make any claims related to specific genre-based revisions, and hence cannot create concrete insights for genre-based instruction. In addition, the tagset is focused on revision in text, and does not consist of features specific to visual components, such as images or tables (cf. Leijten et al., 2014). However, we tried to keep the current tagset as context-independent as possible. In this way, the tagset can be used in every context and, where necessary, be tailored to specific genres, languages, or tasks.

Another limitation is that only keystroke logging, screen replays, and eye tracking were used for the manual and automatic annotation of the features. This made it more difficult to manually annotate several features, for example for orientation or domain, because the intention of the writer was not always known. It would be interesting for future work to triangulate the data with data from thinking-aloud or retrospective interviews. This could reveal if and in what cases these time-intensive methods are preferred over automated data collection such as keystroke logging and eye tracking.

5.4.2 Implications for research and future work

The tagset has several implications for research and future work. First of all, the comprehensive classification of revision makes it possible to study revisions in more depth. Until now, revisions have often been operationalized rather mechanically, as the number of backspaces or deletions (see e.g., Zhu et al., 2019). Although this is a quick approach, it does not allow for a thorough understanding of what types of revisions are made. For example, the tagset could be used to identify how the types of revisions differ across groups of writers (e.g., L1 versus L2, novice versus expert; cf. Stevenson et al., 2006), different tasks (e.g., genres, timed versus non-timed, cognitive load; cf. Révész et al., 2017), or different phases in the writing product (e.g., begin, middle, end; cf. Barkaoui, 2016). An example case study on how this could be achieved with the current tagset can be found in Conijn, Dux Speltz, et al. (2020).

In addition, as the tagset is both product-oriented as well as process-oriented, the tagset makes it possible to better understand the process of revision and its effect on the writing process. Pattern mining or clustering can be used to find sequences or clusters of similar revisions. In this way, we could further investigate the sequencing of revisions, the causes and the effects of the revisions, and how this differs for different types of revisions. This can eventually be used for a better alignment of keystroke logging with writing processes, as different revisions 'behave' differently (Galbraith & Baaijen, 2019).

It is important to note here that it is not necessarily essential to include the full tagset in future work. Rather, the tagset is a basis of which the relevant properties, depending on the research question at hand, should be selected. For example, for more process-oriented analysis, the temporal and spatial location of the revision in the writing process would be important (cf. Breetvelt et al., 1994). In addition, the tagset allows for better informed choices and transparent descriptions of which features are included (and which are not).

Although for these future studies several features still need to be manually annotated, the richness of the tagset also allows for classification of the manual features using machine learning (also see Chapter 6). In this way, the manual features could be approximated automatically. Some studies already tried to build classification models for some of the manually annotated features. For example, Xue & Hwa (2014) built a classifier to identify the sequencing of revisions: an indication of whether consecutive revisions belonged to the same mistake. For this, they included the following features: number of words between

edits, change of tense, change of word order, change in same word set, edit distance of revised words, original word in dictionary, original and revised word same part-of-speech, and original and revised word both prepositions. Daxenberger & Gurevych (2013) tried to automatically classify the orientation and action features manually annotated in previous work (Daxenberger & Gurevych, 2012). Classifiers were trained to classify the labels for each of the 21 categories in the corpus, using character and word n-grams with $n = 1$, 2, and 3. Likewise, F. Zhang & Litman (2015) also classified manually annotated orientation labels. As features, they included unigrams, spatial location of the revision (e.g., first sentence), textual properties, such as edit distance, named entity, and discourse marker, and language properties, such as part-of-speech, spelling, and grammar mistakes. All these studies showed that their machine learning algorithms beat the baseline classifier and hence that we could, at least to some extent, automatically classify the labels. However, the classification for fine-grained classes, such as distinguishing between a surface change within a word and a conventional change, were still found to be difficult (F. Zhang & Litman, 2015). The current tagset provides additional features that could be used to further improve on these automatic classification algorithms.

5.4.3 Implications for educational practice

For educational practice, the tagset can be used to provide a more detailed overview of revisions. This could be done by, for example, by visualizing the revision properties in the form of a dashboard (also see Chapter 7). Dashboards are tools that provide an overview of students' tracked learning activities to promote awareness and reflection (Verbert et al., 2014). Accordingly, such a dashboard can be used by students to reflect on their revision process, or by teachers to inform their feedback and instruction on the revision process. A small case study already showed that such visualizations on students' writing processes can be useful to discuss and reflect on the writing process with students (see e.g., Vandermeulen et al., 2020). In addition, these visualizations may be used to adjust students' (implicit) conceptions about revision and revision processes, which are argued to differ between teacher and student, and novice and expert writers (Flower et al., 1986). Lastly, the dashboard can be implemented as an intervention targeted at improving revision strategies. The tagset can then in turn be used for a more detailed evaluation of the impact of the intervention on the types and timing of the revision.

5.5 Conclusion

To conclude, in this chapter we presented a comprehensive tagset of revisions, including ten product and process-oriented properties of revisions: processing, trigger, orientation, evaluation, action, linguistic domain, spatial location, temporal location, duration, and sequencing. We have shown that many variables related to these properties can be automatically extracted using information from the keystroke log. In addition, replays of the keystroke data combined with eye tracking could be used for manual annotation of the features that cannot be extracted automatically. These manual features could eventually be classified using machine learning techniques. In this way, the tagset could be implemented within a fully automated tool to provide detailed information on revisions. This allows for a scalable approach to analyzing revision in writing in depth.

6

Building a process-based model of typographic error revisions

Adapted from: Conijn, R., van Zaanen, M., Leijten, M., & Van Waes, L. (2019). How to typo? Building a process-based model of typographic error revisions. *Journal of Writing Analytics*, *3*, 69–95.

In the previous chapter we showed the large variety of types of revisions. Intuitively, one of the most negligible type of revisions is the revision of typographic errors (slips of the finger). However, these revisions can have a large influence on the writing process, and hence also on the analysis of the writing process. On the one hand, these types of revisions are low-level, and hence less important, so it might be advisable to ignore them for certain research questions. On the other hand, it is important to identify these revisions, as they can (unwillingly) break the flow in writing. Therefore, in this chapter we aim to build a process-based model of typographic errors and their revisions. First, we characterize typographic errors and their revisions based on temporal and bigram properties extracted from keystroke data. Thereafter, we train a process-based model on typographic error revisions on a copy task dataset to automatically identify these revisions. Finally, this model is evaluated in a more natural setting: a regular (source-based) writing task. Results show that it is possible to identify typographic errors using keystroke data only, especially in a copy task. Yet, the models tested on the source-based writing task still lead to a high number of false positives. To conclude, using these models, a more nuanced analysis of fluency and revision in writing can be performed.

Acknowledgements. I would like to thank Tineke Conijn for her assistance in annotating the dataset.

6.1 Introduction

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. First, these revisions 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, Jacovina, et al., 2016; Roscoe et al., 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. Second, 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 et al., 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. 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 instruction on revisions in multiple settings. This allows for evidence-based practices within writing instruction (Graham, 2019). Identification of revisions can 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).

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 chapter, 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 (Lindgren & Sullivan, 2019; Leijten & Van Waes, 2013). Several studies have used keystroke data to manually classify revisions into typographic error revisions and other types of revisions (see e.g., New, 1999; Stevenson et al., 2006). However, the automatic analysis of typographic error revisions using process data has received little scholarly attention.

6.1.1 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 et al., 2007). An early review of typing errors already showed the variety of error categorizations that were made (Dvorak et al., 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.

6.1.2 Prediction in writing process versus writing product

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 is 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 et al., 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 the 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.

6.1.3 Prediction of typing errors versus typographic errors

Second, previous studies typically do not distinguish between different types or causes of typing errors (Kano et al., 2007). After all, these studies 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, 1980). Kano et al. (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 & Myers (2006) identified five types of corrected and uncorrected errors: substitution, insertion, omission, no-change, and non-recognition error, 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) defined typographic errors as "slips of the keyboard, i.e., errors that occur despite the writer's knowledge of how they are spelled" (p. 73). Slips can be seen as a human error where the action was not performed as it was intended (Norman, 1981). In Rumelhart & 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.

6.1.4 Prediction of typographic errors in writing process

In this study, typographic error revisions are identified using keystroke data. Keystroke data can provide insight into the writing process (Leijten & Van Waes, 2013; Lindgren &

Sullivan, 2019). However, keystroke analysis has not yet been used to automatically identify all 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 or linguistic errors (errors that break the conventions of written language), such as errors in spelling, grammar, or punctuation (Lindgren & Sullivan, 2006a). Therefore, usually detailed rules or guidelines are 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 may be used for the automatic identification of typo-

graphic errors. Some of these patterns are related to semantic content and pronunciation, which may be hard to extract when the intended text is unknown. Other patterns may 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 may influence the probability of a typographic error.

In this chapter, 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. This is done in three steps. First, we identify what process-based features might be indicative of typographic errors and their revision. Second, we determine how these features could be used to classify typographic error revisions in a copy task. Lastly, we identify whether this model also transfers to a more natural writing task.

6.2 Method

For the three steps, three analyses were conducted using two different datasets; a copy task dataset and a source-based writing task dataset. First, we characterized typographic errors and revisions in a dataset from a carefully manipulated copy task (see Section 6.2.3). In a copy task, all errors can be considered typographic errors since the correct spelling and grammar are 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 (see Section 6.2.4). 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 (see Section 6.2.5). In the following, we first discuss the two datasets collected and the cleaning and transformation of these data, followed by the three analyses conducted.

6.2.1 Copy task dataset

The copy task data has been collected from the Dutch copy task in Inputlog (Van Waes et al., 2019). 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%).

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 6.1. The participants were asked to write each word combination seven times. In total, there were 59,423 attempts of the word combinations, consisting of 1,445,314 characters. After every keystroke in every attempt, 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.

After data cleaning, typographic errors and revisions were coded in the keystroke data file. Typographic errors were extracted largely following the procedure as described by Wobbrock & 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 be-

Table 6.1: Bigram properties of the word combinations in the Dutch copy task in Inputlog (Van Waes et al., 2019)

		Number of bigrams			
Word combinations	Length	High freq.	Low freq.	Adjacent	Repetitive
vier mogelijke verbanden	23	0	20	7	0
drie belangrijke kinderen	23	0	19	7	0
vier duidelijke manieren	22	0	19	8	0
een chaotische cowboy	19	4	12	0	1

	Number of bigrams			
Word combinations	Left–Left	Left–Right	Right–Left	Right–Right
vier mogelijke verbanden	4	5	4	4
drie belangrijke kinderen	4	4	3	4
vier duidelijke manieren	4	5	5	5
een chaotische cowboy	2	3	2	0

tween 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.

6.2.2 Source-based writing task dataset

The source-based writing task (synthesis) dataset was a subset of the dataset collected by Leijten and colleagues (Leijten, Van Horenbeeck, & Van Waes, 2019; Leijten, Van Waes, et al., 2019). All the participants in this subset were also in the copy task dataset. The participants were asked to write a text in Dutch 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).

6.2.3 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 Section 6.1.1). 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.

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.

Bigram properties. Bigram properties were extracted from character bigrams, hereafter referred to as bigrams. Bigram properties were extracted from the typed bigram in which the typographic error was made, the expected bigram (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 6.2. 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 6.2: Examples of possible typographic error types with the typed and expected text and bigrams

	Word		Character bigram		
Error type	typed	expected	typed	expected	swapped
Deletion	moglijke	mogelijke	gl	ge	l→e
Insertion	moogelijke	mogelijke	oo	og	o→g
Substitution	mofelijke	mogelijke	of	og	f→g
Transposition	mogeiljke	mogelijke	ei	el	i→l

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 et al. (2017), using the CELEX Lexical Database of the Dutch Centre for Lexical Information (Baayen et al., 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).

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.

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 charac-

ters 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.

6.2.4 Analysis 2: Automatic classification of typographic error 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.

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 6.2. As in the characterization, bigram frequency, hand combination, key adjacency, and repetitiveness were extracted.

Timings. The timing of the keystrokes surrounding the typographic error were calculated using the IKI of these keystrokes. Again, the timings were calculated for the IKI of the current keystroke and the four preceding keystrokes, as well as the IKI of the five following keystrokes.

Participant. Participant ID was included, as typographic errors are made differently across typists. Hence, this will probably result in a better model. However, this also means that the model will not generalize to new students. Accordingly, when predicting typographic errors for a new student, copy task data of this student would be necessary.

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, 2019) 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 (dummy coding) was used to transform the categorical bigram properties into binary features. Centering and scaling were applied to the timing features.

Two models were run, using: (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, 2019), 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:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{6.1}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{6.2}$$

$$\text{F-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{6.3}$$

6.2.5 Analysis 3: Automatic classification of typographic error 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 (see Section 6.2.4). 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. By 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 *hoise*), 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.

Finally, false positives and false negatives were analyzed to gain insight into possible points of improvements for the model. To analyze the errors, the keystroke data surrounding the wrongly identified typographic revision or missed typographic revision were 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 errors that could be easily solved, should be prioritized in the next iteration of the model.

Table 6.3: 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

6.3 Results

6.3.1 Characterization of typographic error revisions

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 6.3 presents an overview of how often each type of error was revised. Substitutions were revised most, followed by transpositions. Deletions were least often revised.

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 6.1). 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.

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).

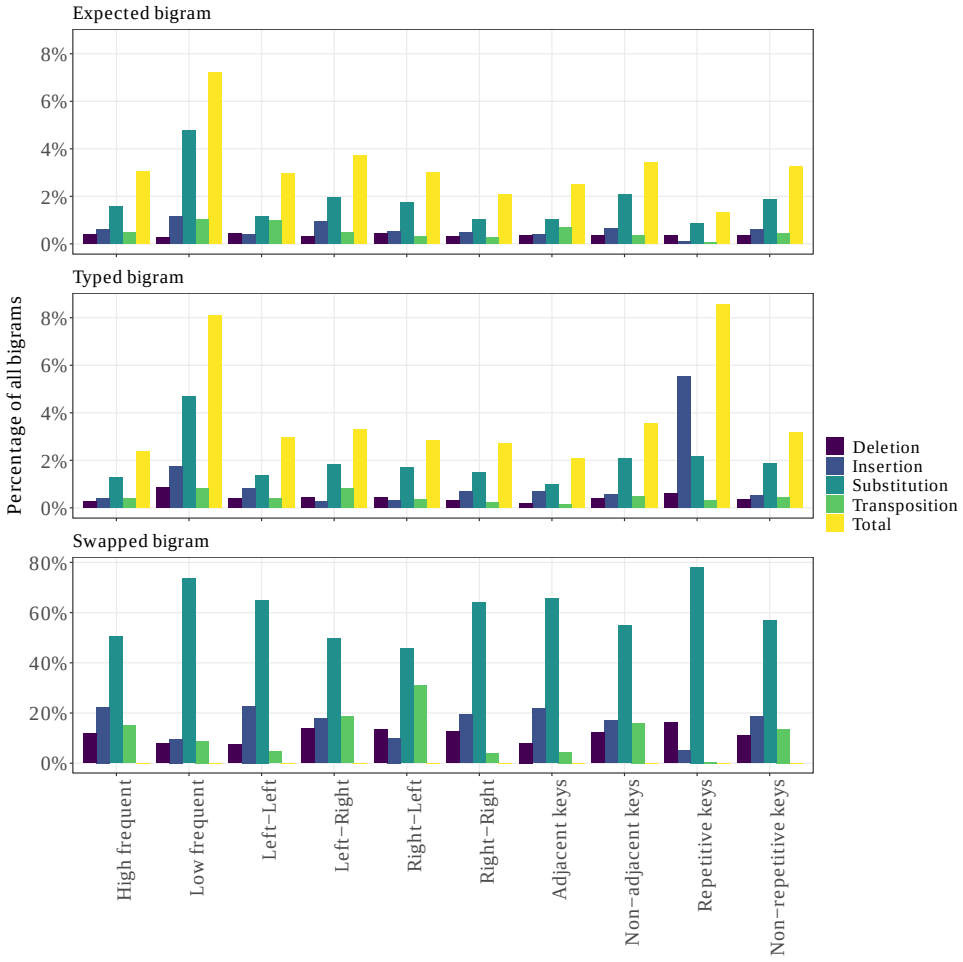

Figure 6.1: The percentages of errors within each bigram feature for expected and typed bigrams.

The bigram properties for the different types of error in the swapped bigrams showed a somewhat more distinctive pattern (see Figure 6.1). 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*).

Timing. The timings of the keystrokes around the typographic error showed a relatively clear pattern for the error (see Figure 6.2). The IKI of a revised typographic error was on average 46 ms longer than the mean IKI of the participant. Thus, typists slow down when making a typographic error. Yet, the variance was large (SD = 170 ms), 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 37 ms longer than the mean IKI, at the keystroke of the typographic error. Again, the variance was large (SD = 203 ms).

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 = 59 ms, SD = 276 ms 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 = 201 ms, SD = 251 ms), 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.

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-change revisions (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 by one keystroke (25%), two keystrokes (9%), three keystrokes (5%), or more than three keystrokes (5%). Sometimes the revision was even delayed by more than 10 keystrokes (0.1%). The latter indicates that the revision was made over word boundaries. Lastly, in the current dataset, it was relatively uncommon to revise more keys than necessary: there

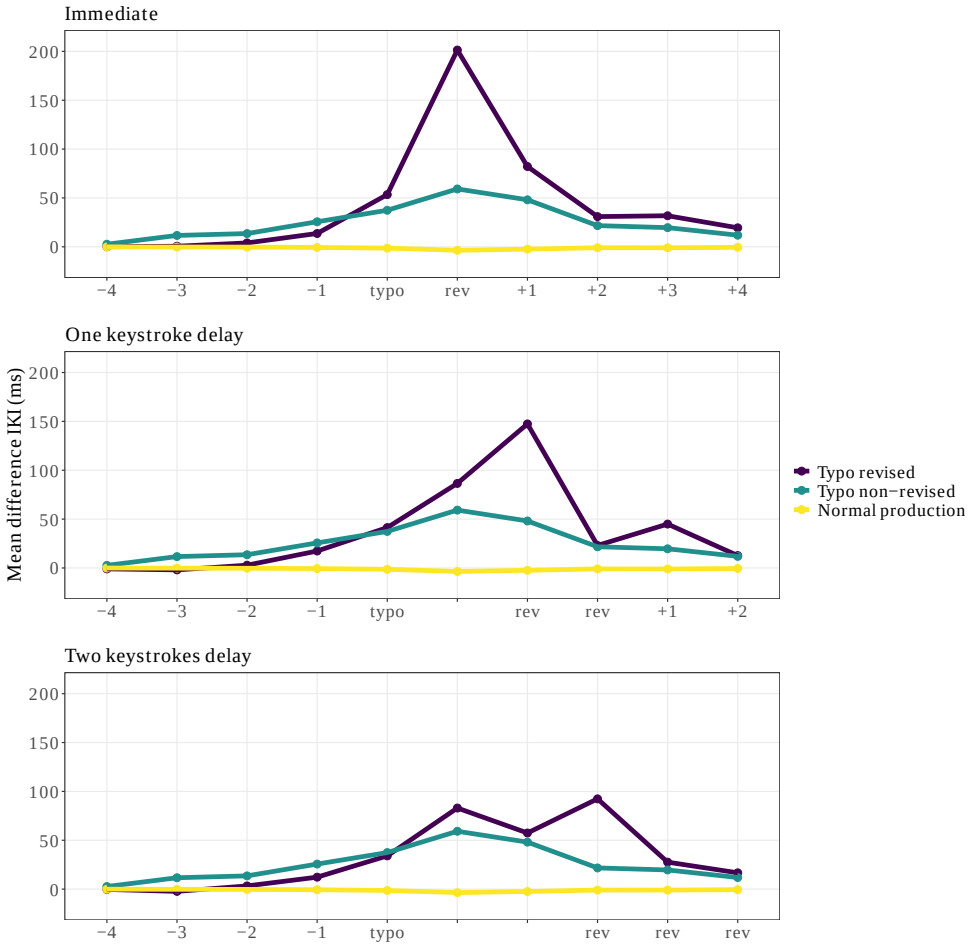

Figure 6.2: The interkeystroke interval (IKI) before and after (revised) typographic errors, compared to normal production.

was a revision overflow of one letter in only 4% of the revisions, 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

Table 6.4: Performance of the prediction of typographic errors in the copy task dataset

	All features, *M*(*SD*)			Features up to error, *M*(*SD*)		
Model	Precision	Recall	F-score	Precision	Recall	F-score
Random forest	**.86(.02)**	.86(.01)	**.86(.01)**	**.75(.02)**	**.71(.04)**	**.73(.02)**
Support vector machine	.82(.02)	**.87(.03)**	.84(.02)	.67(.03)	.67(.04)	.67(.03)
Logistic regression	.75(.01)	.79(.03)	.77(.02)	.66(.02)	.66(.04)	.66(.02)
Naive Bayes	.64(.03)	.78(.03)	.70(.03)	.62(.01)	.52(.03)	.57(.02)

Note. Majority class baseline accuracy is .5. Bold values indicate highest performance.

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 following we classify typographic errors using the process variables described above.

6.3.2 Automatic classification of typographic error 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 6.4). Random forests were overall the best model, with a precision of 0.75, recall of 0.71, and an F-score of 0.73. By 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, recall, and F-score of 0.86. The support vector machine was only slightly better in recall, with a recall of 0.87. Thus, process-based data can be used to predict a typographic error with a relatively high performance.

6.3.3 Automatic classification of typographic error 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,

Table 6.5: Number of revisions predicted as typographic revision in the source-based writing task dataset

Model	All features	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 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 6.5). 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.

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 6.6). The models including all features showed better results. Again, random forests proved to be the best model (accuracy = 59%), but 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.

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-change revisions, where character(s) were replaced by the same character(s). Second, some of the false

Table 6.6: Accuracy of the prediction of typographic revision in the source-based writing task dataset

	All features			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. Bold values indicate highest performance.

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 were found 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 word initial location or when the error was only revised after the word was finished (including the space after the word).

6.4 Discussion

In this chapter 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 of-

ten low-frequent bigrams than high-frequent bigrams. *Swapped* bigrams, often presented in confusion matrices of typing errors (see e.g., Kernighan et al., 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 & 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).

For the second analysis, a process-based model was built on typographic errors in a copy task, based on insights gained from the characterization of typographic errors. 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).

Lastly, these models were 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.

6.4.1 Limitations and future work

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 often revised with delays of several 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. 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.

Finally, future work should identify how the model could be further improved. For several types of typographic error revisions, such as capitalization, and transposition of an adjacent key, a rule-based approach might be sufficient (Kim, 1996). To extract more complex typographic error revisions, 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-change revisions were the most common false positives, indicating that the largest improvement gain might be reached when this error is addressed. No-change revisions 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-change revisions 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.

6.5 Conclusion

To conclude, this chapter provides 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 very accurate yet, 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, to identify which types of revision students focus on the most.

7

Human-centered design of a dashboard on students' revisions

Adapted from: Conijn, R., Van Waes, L., & van Zaanen, M. (2020). Human-centered design of a dashboard on students' revisions during writing. In *Conference proceedings of the 14th European Conference on Technology Enhanced Learning, EC-TEL* (pp. 1–15). https://doi.org/10.1007/978-3-030-57717-9_3

The last two chapters showed how we could extract information (partly automated) on students' revision processes. This data could be visualized in the form of a learning dashboard. Learning dashboards are often used to provide teachers with insight into students' learning processes. However, simply providing teachers with data on students' learning processes is not necessarily beneficial for improving learning and teaching; the data need to be *actionable*. Recently, human-centered learning analytics has been suggested as a solution to realize more effective and actionable dashboards. Accordingly, the current chapter aims to design an *interpretable* and *actionable* dashboard to provide insight into students' revision processes, using a human-centered approach. The design consists of three iterative steps. First, visualizations on students' revision processes, created from keystroke data, were evaluated with writing researchers. Second, the updated visualizations were used to co-design a paper prototype of the dashboard within a focus group session with teachers of academic writing. Finally, the paper prototype was transformed into a digital prototype and evaluated by other teachers in individual user tests combined with interviews. The results showed that this approach was useful for designing an interpretable dashboard with envisioned actions, which could be further tested within real-world classroom settings.

7.1 Introduction

In most classroom environments, teachers are not able to systematically monitor students' writing processes. Teachers have access to the final written products of students (and sometimes intermediate products), but they typically have limited information on how students create these final products. Hence, little is known on where and when students struggle during their writing.

Writing analytics can be used to provide teachers with more insight into students' writing processes. Writing analytics focuses on "the measurement and analysis of written texts for the purpose of understanding writing processes and products, in their educational context, and improving the learning and teaching of writing" (Buckingham Shum et al., 2016, p. 481). Keystroke logging is often used as a tool in writing analytics to gain insight into students' writing and revision processes (Lindgren et al., 2019). Real-time keystroke data offer the potential for automatic extraction of important diagnostic information on students' writing processes, making it possible to provide a more precise identification of writing difficulties (see e.g., Likens et al., 2017).

Yet, the diagnostic information from fine-grained keystroke data are often not directly intuitive for educational stakeholders. A solution would be to provide the data in the form of a *dashboard*. A learning dashboard is a tool that provides a visual overview of students' tracked learning activities (e.g., time spent on quizzes, number of learning sessions), to promote awareness and reflection (Verbert et al., 2014). Dashboards can be employed by teachers, students, or other educational stakeholders to review and analyze data on students' learning processes (Verbert et al., 2013, 2014). Teacher-facing dashboards are specifically aimed at providing teachers with information to improve their teaching and students' learning. These dashboards have been proven to be useful and effective in improving posttest scores and engagement (Verbert et al., 2014).

There is limited research on designing dashboards for visualizing students' writing processes. To our knowledge, only two studies address this issue. One study designed a dashboard on collaborative writing (Olson et al., 2017; Yim et al., 2017), in which DocuViz (a Chrome plugin) is used to visualize how a document grows over time, who contributed, and at which times and locations people contributed. Another study describes the process report in Inputlog, which shows pausing behavior, revision behavior, source interaction, and fluency (Vandermeulen et al., 2020). This report is mostly textual, but also includes

two visualizations: the progress graph and the fluency graph, detailing the text production and the fluency over time. Hence it has some similarities with a dashboard.

One concern with dashboards, and learning analytics in general, is that simply providing teachers with data on students' learning processes is not necessarily beneficial (Wise & Jung, 2019). It has been argued that for the learning analytics to be effective in improving learning and teaching, the data need to be transformed into *actionable* information (Conde & Hernández-García, 2015). It is not always clear how to act upon insights obtained from tracked learning activities to improve learning and teaching. For example, even though expert revisers might revise more or make more extensive revisions; simply asking novices to revise more rarely results in higher writing quality (cf. Flower et al., 1986).

Recently, human-centered learning analytics has been suggested as a solution to realize more effective and actionable learning analytics (Buckingham Shum et al., 2019). Within a human-centered approach, the functionality and design of the system is defined by the actual users of the system, rather than by the developers or researchers (Buckingham Shum et al., 2019; Giacomin, 2014). Accordingly, the design is more likely to account for all the needs, desires, and experiences of the relevant stakeholders (Buckingham Shum et al., 2019; Giacomin, 2014). As teachers are the main users of teacher-facing dashboards, they have a central role in the design of the dashboard in the current study. By including the teachers in the design process, the alignment of the learning analytics with the learning design of the course or module could be enhanced (Lockyer et al., 2013). When the information provided matches how teachers teach their courses, it is easier for teachers to integrate the analytics into their daily teaching practices and to relate the information to their teaching concerns (Wise & Jung, 2019). Accordingly, the actions upon these analytics will increase and hence the benefits will be higher.

Several frameworks and models have been proposed for human-centered approaches to design learning dashboards (Dollinger et al., 2019). These approaches have been used for the design of dashboards on for example collaborative learning (Martinez-Maldonado et al., 2013), student engagement in courses (Dollinger et al., 2019), general course activities (Wise & Jung, 2019), and reading (Tan et al., 2017). In the field of writing, human-centered approaches have been advocated as well (Buckingham Shum et al., 2016; Cotos, 2015; Knight et al., 2017). In addition, for the design of writing tools in general, it has been shown that these tools are less efficient and used less if they are not aligned with the

classroom activities (Link et al., 2014). By contrast, when writing tools are tuned to the educational context, they are more positively perceived by students, resulting in a higher adoption (Shibani et al., 2019). However, evidence of the use of these human-centered design approaches for the design of dashboards for teaching writing is limited.

Given the limited research on teacher-facing dashboards on students' writing processes and the paucity of research on human-centered approaches for designing writing analytics dashboards, the current study describes a human-centered approach to inform the design of a teacher-facing dashboard on students' writing processes. Specifically, this chapter aims to design a writing analytics dashboard which can transform data on students' writing processes into both *interpretable* and *actionable* information.

7.1.1 Human-centered approaches in dashboard design

Several human-centered approaches have been proposed for the design of learning analytics tools and dashboards (Dollinger et al., 2019), sometimes using different terminology, such as user-centered design, human-centered design, participatory design, co-design, and co-creation (for an overview of all these terms and there definitions, see Dollinger et al., 2019). In the current study, we employ two frameworks that specifically focus on the human-centered design of learning analytics dashboards: an iterative workflow and a model for teacher inquiry.

Martinez-Maldonado et al. (2015) developed an iterative workflow for designing and deploying awareness tools, called Learning Awareness Tools – User eXperience (LATUX). Awareness tools are defined as tools that provide teachers with an enhanced level of awareness on students' behavior within their learning environment. Hence, these could be seen as learning analytics dashboards. The LATUX workflow is based on four principles: (1) the possibilities of the data; (2) how the context influences the stakeholders' design needs; (3) the holistic design and evaluation of the tool, within its context; and (4) the evaluation based on both the data and the actions upon these data (Martinez-Maldonado et al., 2015). These principles are integrated in five iterative design stages. The first stage includes the problem statement and requirements identification. In the second stage a low-fidelity prototype is created, followed by a high-fidelity prototype in the third stage. In the fourth stage, the prototype is evaluated in pilot classroom studies and in the last stage the tool is evaluated in the wild, the classroom itself, to evaluate the longer-term impact on learning and instruction.

Teacher inquiry is of key importance in each of these stages. In this context, Wise & Jung (2019) developed a model for actionable learning analytics to inquire into how teachers are using learning dashboards. This model describes the actions involved in teachers' analytics use, which are divided into the sense-making of the data and the pedagogical response based upon the data. The sense-making consists of how teachers ask questions about the data displayed in the dashboard, where they start looking in the dashboard, which reference points they use, how they interpret these data, and which sources they use to explain the data. The pedagogical response consists of how teachers (intend to) act upon these data, how they align this with their learning design, what impact they expect, and how they measure this impact.

The design of the writing analytics dashboard in this chapter will follow the LATUX framework (Martinez-Maldonado et al., 2015) for the design of a writing analytics dashboard. The first stage (problem statement and requirement identification) can be found in Chapter 2. In this chapter, we report on stage 2 and stage 3. In both stages we apply the model for actionable learning analytics (Wise & Jung, 2019) as a guide to identify teachers' sense-making of and pedagogical response to the dashboard.

7.1.2 The revision dashboard

The dashboard designed in the current chapter is focused on displaying information on students' tracked writing processes, for two reasons: (1) students' self-report on their writing processes is unreliable and (2) analysis of the writing product does not always provide insight into the when, where, and why students struggled (Ranalli et al., 2018b). The dashboard is specifically focused on revisions, because revisions play an important role in writing (Fitzgerald, 1987; Flower & Hayes, 1980; Scardamalia & Bereiter, 1983), but are not visible in the written product. Revisions provide an indication of issues or point of improvements a writer identified in their text (Flower et al., 1986). In addition, revisions can influence the writing quality, the writing process (e.g., disrupting the flow), and the writers' knowledge about the topic or about writing itself (Barkaoui, 2016; Fitzgerald, 1987). Accordingly, revisions play an important role in teaching of writing. Therefore, we designed a teacher-facing revision dashboard to provide teachers with insight into students' revision processes. Data for the revision dashboard is obtained via keystroke logging. In addition, we use the revision tagset as proposed in Chapter 5, to transform the raw keystroke data into visualizations on different properties of revision.

7.2 Method

A human-centered design approach was chosen to design the revision dashboard. The qualitative approach described in this chapter is divided into three iterative design steps. The first two steps align with stage 2 from the LATUX framework (both low-fidelity prototyping) and the last step with stage 3 (high-fidelity prototyping; Martinez-Maldonado et al., 2015). First, visualizations on students' revision processes were created from keystroke data, and evaluated with writing researchers in round table sessions (N = 13). Second, the updated visualizations were used to co-design a paper prototype of the dashboard during a focus group session with writing teachers (N = 4). Third, the paper prototype was transformed into a digital prototype, and evaluated by writing teachers in individual user tests combined with interviews (N = 6).

The full study was approved by the school-level research ethics and data management committee. All participants provided informed consent. None of the participants participated in more than one step.

7.2.1 Step 1: Creating visualizations on revisions

For the first step, three short round table sessions were held in three groups of 4–5 people each (N = 13). The goal was to get quick evaluations on a variety of different visualizations that were made on students' revision processes. Therefore, writing researchers were chosen as participants, as they generally have a high data literacy, and hence can provide feedback on data visualizations without too much explanation. All participants were attendants of a writing research meeting who were willing to participate and who had at least two years of experience in writing research.

The materials included twenty visualizations from students' revision processes, see Appendix C (for an example see Figure 7.1. These visualizations were created in R based on an open source dataset on students' revisions (Conijn, Dux Speltz, et al., 2020). This dataset is annotated on eight properties of revisions: orientation, evaluation, action, linguistic domain, spatial location, temporal location, duration, and sequencing, based on the revision tagset from Chapter 5. Each visualization was made on one or a combination of two or three of these properties. All visualizations were printed on paper.

First, the different properties of revisions were explained. Thereafter, the visualizations were semi-randomly spread out on a large table (visualizations from the same properties of

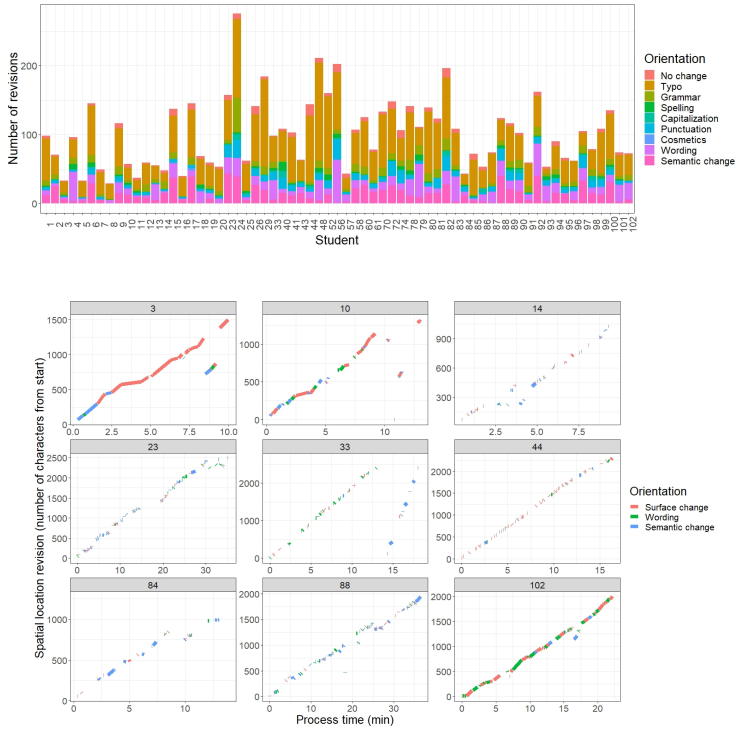

Figure 7.1: Two visualizations of the revision properties for step 1, top: evaluation; bottom: orientation, spatial and temporal location.

the revision tagset were kept together). The participants were asked to individually evaluate the visualizations, using colored post-its. Red post-its indicated aspects they disliked or points of improvement; green post-its indicated things they liked; and yellow post-its were used for questions. In addition, the participants were asked to individually vote on (at most) three visualizations they liked most, using colored dot votes. The participants were not allowed to talk during the evaluation. After each group finished, all votes and post-its were removed and the visualizations were reordered. In total, each group received 15 minutes to evaluate the visualizations.

As analysis, the post-its describing similar themes were clustered. The comments and votes were used to improve the visualizations and to select a subset of these improved visualizations for step 2 (see Section 7.3.1).

7.2.2 Step 2: Creating a paper dashboard of revisions

For the second step, a focus group session was held with four academic writing teachers, the intended users of the dashboard. The goal was to co-design an interpretable and actionable paper prototype of a teacher-facing revision dashboard. In addition, we aimed to determine how this dashboard could be used in the teachers' learning and teaching practices. The teachers were recruited by email via the university's language center. All teachers had at least seven years of experience in teaching academic writing in Dutch and English within classes ranging from 40 to 120 students (three teachers) and/or individual coaching (three teachers).

The materials included nine improved visualizations from step 1, see Appendix D (for an example, see Figure 7.2, p. 172) and four visualizations related to additional information on the writing process: the writing assignment, the replay of the writing process, the final text, and the total time spent on the assignment. All visualizations were printed on paper.

During the focus group, the participants first received a brief introduction on writing processes and dashboards. The participants were asked to discuss properties of dashboards they liked and disliked (15 minutes). Thereafter, the participants were asked to vote on at most five visualizations they would like to have in the interpretable and actionable dashboard, using colored dot votes. Questions on specific visualizations could be added with yellow post-its. All the visualizations were stuck to the wall, requiring the participants to walk around (5 minutes). After voting, the visualizations with the most votes were discussed and the post-it questions were answered by the moderator (10 minutes). The visualizations with the most votes were then pasted onto an empty sheet on the table, to form the paper dashboard. Finally, the participants were asked to discuss, based on this paper prototype (1) what actions they would take based upon this dashboard, and (2) what visualizations or information needed to be changed or added to better inform those actions (50 minutes). The focus group took 90 minutes in total.

During the focus groups, audio was recorded. The audio was transcribed and coded using NVivo 12. The properties of dashboards that the participants liked and disliked were clustered into similar themes together with the participants during the focus group itself, and written on a flip-over sheet. The remainder of the transcript was coded based on the topic list created from the teacher inquiry model for actionable learning analytics from Wise & Jung (2019), see Table 7.1 (first column). The codes were analyzed using thematic

Table 7.1: Topic list steps 2 and 3

Topic	Example question
Asking questions	What do you look for in the dashboard?
	Are there specific aspects you are looking for in the dashboard?
Orientation	Where do you start looking in the dashboard?
Reference point	What is your reference point?
	Is a reference point necessary here? Why?
	How do you compare these visualizations/graphs?
Interpretation	How do you interpret this?
	Which conclusions do you draw here?
	Why is this (un)clear?
Explanation	How do you explain this?
	Which sources/visualizations do you need to explain this? Why?
Action	How would you act upon this?
	What do you as a teacher need to do?
	What do the student(s) need to do?
Alignment	How do you align this in your course?
	Where do you implement this in your course design?
	(How) do you iterate on this over multiple lectures/courses?
Expected impact	What is your envisioned outcome/impact/goal?
	What would be the minimal gain(s)?
Measure impact	How do you measure the impact?
Other comments	Do you have other comments about the dashboard?
	Do you have other comments about your pedagogical response to the dashboard?

analysis. The analysis and the designed materials (paper prototype) were used as input to create the digital prototype for step 3.

7.2.3 Step 3: Evaluating a digital dashboard of revisions

For the third and final step, a series of individual user test interviews was held with six academic writing teachers. The goal was to evaluate the digital prototype of the revision dashboard developed based on the results of step 2, both on interpretability and actionability. The teachers were recruited by email from another university. All teachers had at least

two years of experience in teaching academic writing within classes ranging from 20 to 120 students (five teachers) and/or individual coaching (three teachers). Four participants were already familiar with the progress graph from Inputlog (Leijten & Van Waes, 2013), so they had some prior knowledge on writing process visualizations.

The materials included the digital prototype obtained from step 2, which was shown in the web browser on a personal computer with a full-HD monitor (resolution: 1920 x 1080 pixels).

The interview sessions included thinking-aloud user testing of the digital dashboard, combined with interview questions on the interpretability and actionability of the dashboard. During the interviews, the participants were first shortly introduced on the goal of the interview and the thinking-aloud procedure. After the explanations, the participants were asked to practice thinking-aloud during a specific task (look for the most recent discussion post in a certain course) in a learning management system dashboard. Thereafter, the participants were asked to explore the dashboard and to voice all actions (e.g., button presses), expectations, interpretations of the visualizations, and aspects they found unclear. After the user test, the participants were interviewed on how they would act upon the data in their own course, how they would align the use of the dashboard with their course design, what the expected impact would be, and how they would measure this impact.

During the user test interviews, audio was recorded. As in step 2, the audio was transcribed and coded based on the topic list (see Table 7.1, first column) using NVivo 12. The codes were analyzed using thematic analysis.

7.3 Results

7.3.1 Step 1: Creating visualizations on revisions

The comments and votes on the visualizations of the revision properties in the first step were used to improve the visualizations. The revision properties with at least two votes were used as input for step 2. For five of the revision properties both a frequency-based and a percentage-based graph were shown, resulting in mixed responses from the participants. The graphs with percentages were said to ease finding differences between students. However, especially the participants with a teaching background argued that frequencies were more intuitive to understand for teachers and students. Therefore, we only selected the graphs with displaying frequencies, leaving nine visualizations for step 2.

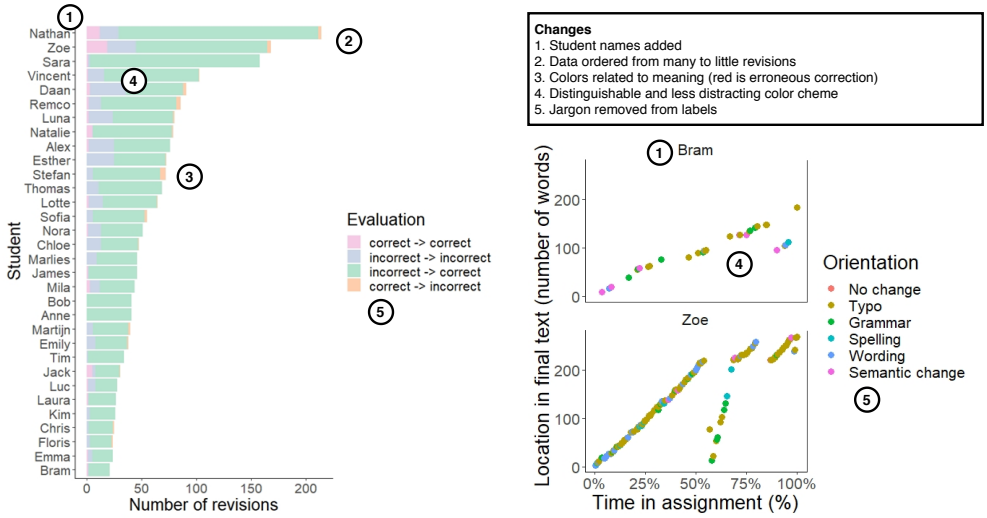

Figure 7.2: Changes in two visualizations of the revision properties after step 1, left: evaluation; right: orientation, spatial and temporal location.

These nine visualizations were improved based on the comments (see Figure 7.2). First, the interpretability of the visualizations was improved based on the visual appeal. The colors were made more distinguishable, where possible related to their meaning (e.g., associating red with errors), and less distracting (e.g., use light colors for dominant categories). The graphs were transformed and the data were ordered from high (top) to low (bottom) frequencies. Additionally, the graphs were improved based on the content. Many participants argued that the labels were sometimes too complex. Accordingly, the labels were changed to reflect student names, and the axis titles were changed to include less jargon. Finally, four visualizations were added to better contextualize the visualizations: the description of the writing assignment, the replay of the writing process, the students' final text, and the total time spent on the assignment.

This first step provided most insight into the design of an *interpretable* dashboard, but some comments related to the *actionability* were already made. The majority of the participants argued that it was hard to directly identify how to use the graphs. One participant mentioned that teachers need to have a goal for using the visualizations. Some of the possible actions mentioned included: goal-setting for students, providing feedback to students, and let students reflect on their writing process (possibly compared to a peer).

7.3.2 Step 2: Creating a paper dashboard of revisions

In the first part of the focus group session, the participants discussed what properties of a dashboard they liked and disliked. The participants argued that the dashboard should give a quick overview. Hence, everything should be clearly organized and visible on one screen, that is, not scrollable. There should be a step-wise approach, where only information that is always relevant is displayed on the homepage, with a possibility to gain more details when clicking on specific visualizations or buttons. The participants liked to have customizable dashboards, to have control over what gets displayed and how this gets displayed. The operation of the dashboard should be simple, intuitive, and easy to learn, especially also for teachers with less technical expertise. Lastly, the dashboard should be attractive. This set the scene for the co-design of the revision dashboard.

In total four visualizations were chosen to be included in the dashboard. One visualization was added during the discussion, when the participants realized they missed the concept of time in the visualizations to fully interpret and act upon the visualizations. The results of the discussion on the paper dashboard are reported per topic below.

1. Asking questions. The participants identified several aspects they would look for in the dashboard. They wanted to get insight into students' writing processes, in relation to effort ("*do they really revise?*"), struggles ("*what are they struggling with?*"), quality ("*how does the process relate to quality?*"), and genres ("*how does the process change for different genres?*"). In addition, one participant mentioned they would like to see whether the concepts covered in class were also addressed within the revisions.

2. Orientation. The orientation, or where to start with in the dashboard was only briefly discussed. The participants argued it should be step-wise, starting from an overview, where it should be possible to go deeper into the data.

3. Reference point. Two reference points were identified to compare the revision process data. First, data could be compared with other students, for example with students who struggled more, with students who received very high or very low grades, or with the class average. In addition, the data could be compared across different texts, for example between drafts, or between different assignments.

4. Interpretation. The participants frequently discussed the interpretation of the revision process visualizations. They mainly focused on the frequency, depth, and distribution of the revisions; the passages students struggled with; and how things might be

interpreted in relation to grades. They discussed how each of the visualizations provided complementary information, but were also wondering about how they could draw generalized conclusions on the visualizations, because "*they are so specific to one persons' process*".

5. Explanation. To explain the visualizations, the participants mentioned four other sources of information that would be necessary. First, the quality of the assignment was needed, to be able to evaluate the revisions. However, some also argued that the quality might not always be useful for explanation, because there is not a single 'best' writing process. Second, access to the final text was necessary, to contextualize the revisions in the actual text that was written (e.g., "*was the revision made in a topic sentence?*"). Third, information on the version of the final text was necessary, as a different revision process might be expected for a first draft, compared to a second draft. Fourth, information on the timing of the revisions was needed and therefore the visualization on the distribution of the revisions over time was added to the dashboard (right graph in Figure 7.2).

6. Action. The participants frequently mentioned they did not see how they should act upon this data. However, throughout the discussions, some ideas emerged. Three possible actions were discussed: encourage students to revise or review more, advise students to focus on deeper revisions, and let students reflect on their processes. In addition, three specific types of scaffolding were discussed: (1) a workshop where students use the tool for multiple assignments, and compare their behavior over time; (2) individual coaching sessions, where the coach goes through the points the student struggled with and, together with the student, finds ways to improve the process; and (3) a class teaching setting, where common 'errors' or struggles are discussed, supported by 'actual' data from the students in class. In addition, the participants mentioned that the tool might be used to measure the effectiveness of their teaching, by identifying whether students are working on the taught constructs.

7. Alignment. The alignment with the course design was only briefly mentioned. Participants mostly wanted to use the tool after the first draft, because then the revision process is most important. The tool should also be used after the grading, because when the student received a good grade, it might not be necessary to look at the process. The participants considered it to be most effective for graduate students, as opposed to undergraduate students, as graduate students already have more insight into their writing and would be more motivated.

8. Expected impact. Five forms of expected impact were determined for the students: get students more motivated, enhance students' awareness and insight into their own writing process, improve their understanding of (the technical aspects of) writing and language, improve the writing process ("*so that they are actually trying to apply what you taught them*"), and improving the writing product. Lastly, an expected impact for the teachers mentioned was to make teaching more fun.

9. Measuring impact. The participants only briefly mentioned some of the expected outcomes needed to be measured, such as the improvement in the writing process over time, and whether students applied what the teacher taught. However, they did not discuss how this could be measured.

10. Other comments. Several other implications of the system were discussed. The participants mentioned that the writing process should not influence their grading. The participants were afraid students might 'game the system', and hence they should not be rewarded for their process. In addition, students' privacy was mentioned, data need to be anonymized, and certain data should not be collected for example if something was produced in the middle of the night. Concerns were expressed on "*yet another system to operate*", hence they preferred to have the dashboard integrated into the university's learning management system. Participants also indicated that it might be very time-intensive to look at the writing process, especially on individual levels. Lastly, the participants were wondering whether the system would still be interesting for teachers to use after five years.

Based on all the results above, the paper dashboard was then transformed into a digital dashboard (see Figures 7.3, 7.4, and 7.5). A demo of the dashboard can be found at `https://rianneconijn.shinyapps.io/Revision_Dashboard`. The dashboard follows a step-wise approach, starting with an overview that could be fleshed out to display the full details. In addition, teachers can filter specific types of revision, to only focus on the types of their interest. To better interpret the data, the participants desired to be able to compare the visualizations between students and between assignments versions. Therefore, the digital dashboard was designed to facilitate these comparisons. Lastly, the participants discussed different teaching contexts: class teaching/workshops and individual coaching, and argued these would require different information in order to be able to act upon the data. Therefore, two tabs were created for each of the teaching contexts, both displaying data preferred in either situation. The digital dashboard was used as input for step 3.

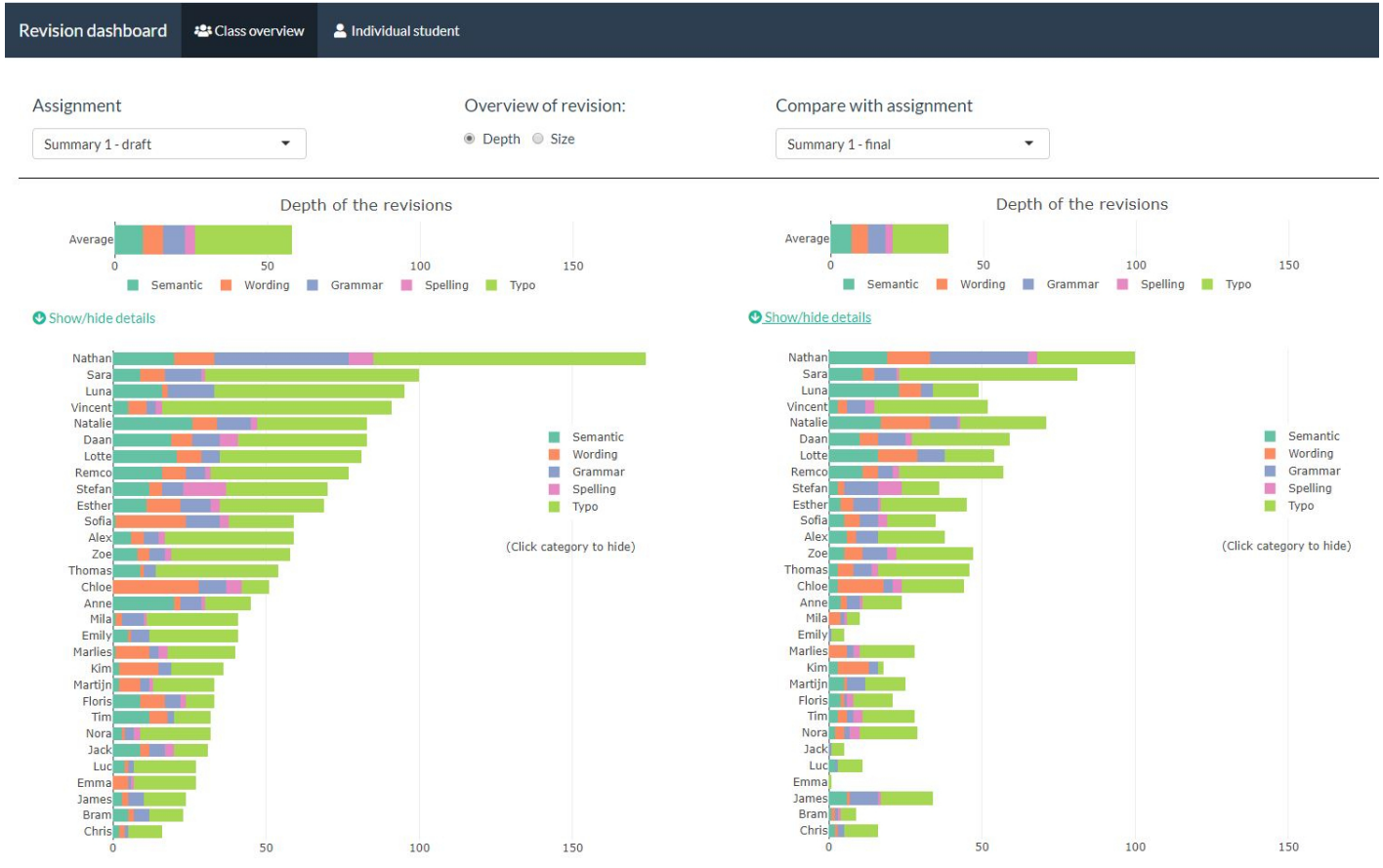

Figure 7.3: Digital dashboard class overview tab, showing the comparison of the different types of revisions made for the whole class, for the draft (left) and final version (right) of a summary task.

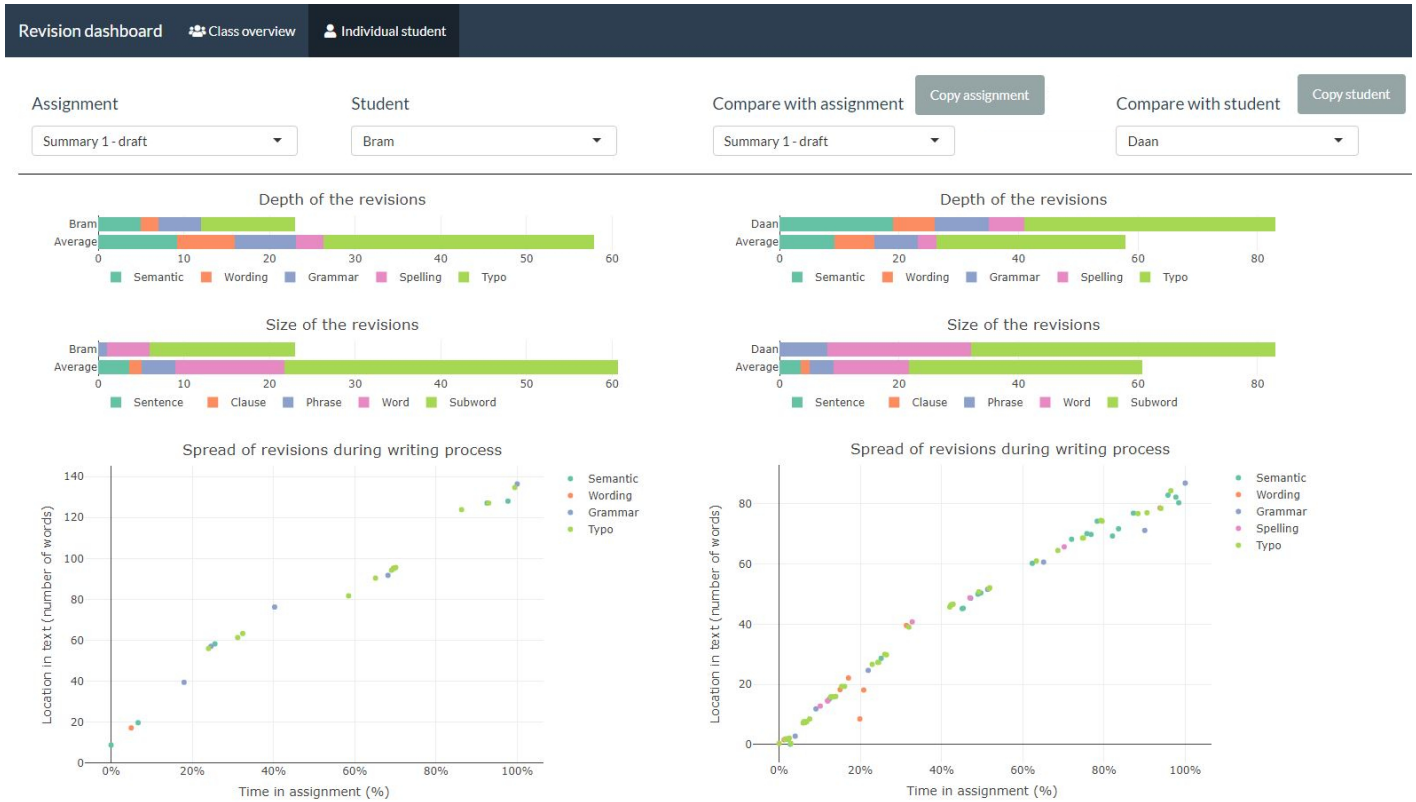

Figure 7.4: Digital dashboard individual student tab top part, showing the comparison of the different types of revisions (top) and the spread of these revisions over time (bottom) for two students: Bram (left) and Daan (right).

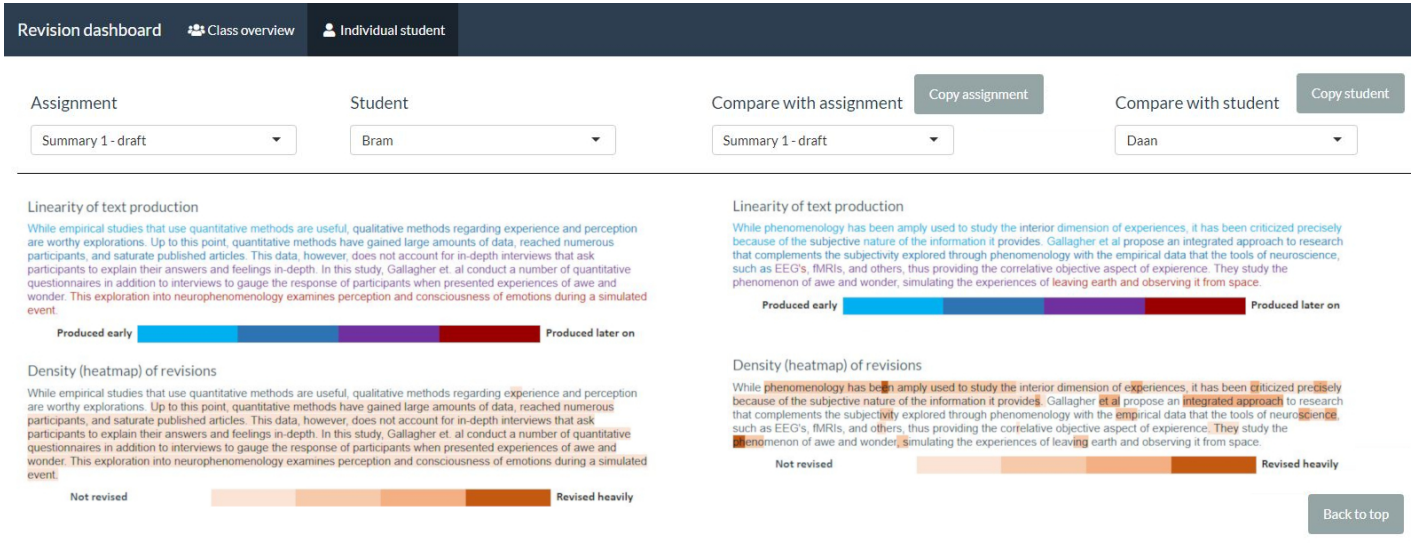

Figure 7.5: Digital dashboard individual student tab bottom part, showing the linearity of the production (top) and the density of the revisions (bottom) mapped onto the written product for two students: Bram (left) and Daan (right).

7.3.3 Step 3: Evaluating a digital dashboard of revisions

The results of the user testing and interview on the digital dashboard are reported per topic below. As the interviews were in Dutch, all quotes are translated.

1. Asking questions. The participants did not start with a specific question; they wanted to explore the dashboard first and asked general questions, such as "*how do students revise?*". More specific questions popped up when the participants explored the data (e.g., "*does everyone show such a linear process?*").

2. Orientation. The design of the dashboard facilitated a step-wise navigation, following the recommendations from step 2. This sometimes resulted in some confusion in the beginning, "*I cannot see anything yet*", however this became clear relatively quickly once they selected an assignment. There were three possible navigation approaches: from class overview towards a specific student that stood out in the class overview; starting from a specific student or assignment because they showed issues; or starting from a random assignment.

3. Reference point. The reference points were the same as in step 2, as the design specifically facilitated these comparisons: compare with other students, previous versions of the same assignment, previous assignments, and the class average.

4. Interpretation. The participants were able to interpret all the graphs, although sometimes clarifying questions were needed. The differences between some of the labels were unclear (e.g., semantic versus wording) and the graph displaying revisions according to the temporal and spatial location (bottom Figure 7.4) took some practice. The interpretations evidently followed the data displayed, and were highly related to the current sample of students within the dashboard. For example, it was noted that the students revised heavily, but mainly focused on minor revisions. In addition, the participants noticed that many students revised at the leading edge of the text, and rarely revised in earlier parts: "*their writing is fairly linear*". Lastly, they noticed how heatmaps showed specific instances of words that were heavily revised and sentences where a student went back to revise their earlier produced text.

5. Explanation. The participants discussed different sources that could aid the explanation of the data. First, the type of assignment was mentioned most, as shorter texts might result in fewer deep revisions, and a final version probably consists of fewer minor (spelling, grammar) revisions compared to a draft version. Second, the quality and final

text was often mentioned, mostly in relation to the possibility of multiple explanations of the revisions. Participants stated that making more revisions does not necessarily result in higher quality and that there is no single 'best' writing process. Hence, they argued it might be hard for students to identify what it means if they make only a few revisions; it needs to be related to their quality or final text to actually draw conclusions. Lastly, three sources were mentioned by only one participant: the number of words so far, the total time spent, and information on students' language background.

6. Action. Contrary to step 2, participants were able to envision a variety of actions based upon the data. Six different actions were discussed: (1) instruction on the writing process, i.e., discussing different strategies and approaches to the writing process; (2) reflection upon students' own processes, to see what could be done differently; (3) discussion on where students struggle; (4) providing advice on what to focus on; (5) measure the effectiveness of teaching; and (6) including information on the writing process into the grading. The same types of scaffolding were discussed as in step 2: individual coaching, workshop, and class teaching. In addition, several participants discussed group work and/or peer-feedback: "*let students discuss their writing processes in groups, to help each other out in understanding the visualizations and identifying how things could be done differently*".

7. Alignment. The participants discussed it would be useful to apply the dashboard in multiple assignments, to be able to compare the data over different assignments. In addition, several suggestions were made to align the dashboard with their current course, such as compare subscores with subgroups of revision and integrate the tool in peer-review, and in the curriculum, to extend on the visualizations over time in the curriculum when students get more experienced with writing and information on the writing process.

8. Expected impact. The impact expected for students was to create awareness and insight into their own revision processes and their struggles, but also in other students' processes, to see there are multiple ways to approach writing. In addition, it was envisioned that students' writing processes would improve, by making more semantic revisions and choosing a more effective revision strategy.

9. Measuring impact. Measuring impact was only discussed superficially, without concrete measures. The participants only discussed measuring the quality of the writing process, which could be done by determining whether students indeed tried a different approach or whether their process evolved over time.

10. Other comments. Lastly, several concerns were discussed. Similarly to step 2, participants were concerned about the privacy of the students, the time needed to use the dashboard in class, and the multiple systems to operate. Four additional implications were mentioned. First, the participants wanted to know how the data were collected, and how accurate they were (transparency). Second, they discussed the learnability of the system: a manual was needed, but, they argued it was easier than they thought. Third, some participants compared the dashboard to other systems, and argued this dashboard was faster than watching a replay of the writing process, and provided more insight compared to the progress graph provided by Inputlog (cf. Leijten & Van Waes, 2013). Lastly, one participant mentioned it might be useful as a tool for research as well.

To conclude, the user test interviews showed that participants were able to interpret the dashboard with only a few clarifying questions needed. This indicated that these steps were successful for creating an interpretable dashboard, which could be further developed in future iterations. In addition, the multiple possible actions envisioned by the teachers, provided evidence for the actionability of the revision dashboard. The results showed suggestions for further improvements of the design of the dashboard, which are described below.

7.4 Discussion

This chapter aimed to design a writing analytics dashboard which can transform data on students' writing processes into both *interpretable* and *actionable* information. This was done within three steps, based upon the LATUX framework (Martinez-Maldonado et al., 2015). First, insights were gained into how the data on different properties of revisions obtained from the revision tagset in Chapter 5 could be best represented, according to writing researchers. Second, based on these visualizations, a paper prototype was co-designed with writing teachers, and evaluated using the teacher inquiry model (Wise & Jung, 2019). This paper prototype was then transformed into a digital dashboard and evaluated with writing teachers, again following the teacher inquiry model.

This process showed several implications for the design of an interpretable and actionable revision dashboard. First, the teacher inquiry model (Wise & Jung, 2019) proved to be useful to get insight into how teachers approached the dashboard, interpreted the visualizations, and envisioned possible actions upon the data. The use cases mentioned within the

envisioned actions are in line with the types of scaffolding found by previous work: whole class scaffolding, targeted scaffolding, and revise course design (Wise & Jung, 2019).

Second, the results showed that for many of the actions, teachers wanted to show the dashboard to the students as well, to discuss and reflect upon their writing process. This indicates that the dashboard should be both teacher and student-facing, and students should be included in future design iterations (Buckingham Shum et al., 2019; Giacomin, 2014).

Third, step 2 and step 3 showed that the teachers preferred to have a step-wise navigation, starting from an overview, before diving into details, and with a possibility to filter out irrelevant information. This reduced the information overload and coincides with previous findings from learning dashboard evaluations (Charleer et al., 2016). However, this also resulted in some confusion in the beginning, as the teachers did not clearly know where to go. In addition, some of the detailed graphs were not clear from the beginning. Therefore, future work should provide a manual with examples or use data storytelling principles (see e.g., Martinez-Maldonado et al., 2020) to better guide the users through the dashboard and improve the interpretability.

Finally, the outcomes also provide further implications for the use of the revision tagset. The results showed that five categories of revision tagset contain the most informative and/or useful for teaching: orientation, linguistic domain, temporal location, spatial location, and sequencing. Hence, these categories might be prioritized for the automatic extraction of these categories using machine learning.

7.4.1 Limitations and future work

The design of the dashboard is not finished. Several concerns were raised for the adoption of the dashboard, resulting in suggestions for further improvements of the design of the dashboard. First, the system needs to be made scalable to allow for large classes. Second the full system (data collection and dashboard) need to be integrated in the existing learning management systems. Third, all participants argued that a manual needs to be provided with examples, especially to explain the differences between all the categories in the different visualizations. Finally, it needs to be further determined how the benefits might be maximized with minimal input from the teachers. These steps should be integrated in further iterations.

This study is limited as it only provides *envisioned* actions upon the dashboard. Further research should explore whether teachers indeed act upon the dashboard. Moreover,

most of the participants had some prior experience with dashboards or information on the writing process, the results might not generalize towards teachers with less experience in this. Hence, the dashboard needs to be evaluated with less experienced teachers as well. Lastly, the interpretability and actionability of the dashboard seemed to largely depend on the data shown in the dashboard. For example, when the data did not show any clear patterns or differences between students, the teachers often had more difficulties interpreting the data and envisioning possible actions. However, it is not always known which differences or patterns might be of interest for teachers. Therefore, it is important to further evaluate the dashboard within the field, using real and contextualized data from the teachers' current students (Holstein et al., 2019).

Further evaluation could be done according to steps four and five from the LATUX framework (Martinez-Maldonado et al., 2015). First, pilot classroom or individual coaching studies need to be considered, followed by longer-term evaluation within the classroom itself. This measures whether the dashboard is actually actionable in practice, and could be used to evaluate the longer-term impact on learning and instruction. In addition, as the learning setting might influence the impact of the dashboard, further evaluations could also be used to identify which actions upon the dashboard are preferred in which situations (cf, Verbert et al., 2014).

7.5 Conclusion

To conclude, this chapter showed the first steps of a human-centered approach into designing a writing analytics dashboard. Using this approach, we envision this will result in a dashboard that can be acted upon, and hence can be effective in improving the learning and teaching of writing.

8

General Discussion

8.1 Summary of findings

In this dissertation, I explored how keystroke logging can be used to gain meaningful insight into students' writing processes. This was done in four steps, with four subquestions, spread over six studies. The discussions on each of the individual studies can be found in the respective chapters. In this general discussion, I first summarize the main findings of these four steps and provide answers to the four subquestions. Next, I describe the overarching limitations and possibilities for future work. Thereafter, I reflect on the use of keystroke logging in the broader context for both writing education as well as writing research. Finally, I conclude.

8.1.1 Identifying stakeholders' needs

When writing, a wide variety of cognitive and behavioral processes may be active concurrently. Keystroke logging can be used to provide an expansive set of features that, at least to some extent, can provide insight into these processes. Accordingly, we first identified which insights into the writing process stakeholders desire, as specified in subquestion 1: *What indicators of students' writing processes are considered desirable, according to multiple stakeholders, for providing feedback on the writing process?*

The five focus groups with educational stakeholders described in Chapter 2 revealed a wide variety of desired indicators of students' writing processes. The indicators covered all main processes in writing: planning, translating, reviewing, and monitoring processes (Flower & Hayes, 1981). Indicators included, for example, information on students' planning strategies, how students use evidence in their writing, the depth of revisions, and students' understanding of the task. For providing automated and personalized feedback, it was considered important to extract behavioral indicators, which could be identified using keystroke logging, such as the speed of writing. However, higher-level cognitive indicators (e.g., critical thinking) as well as behavioral indicators in relation to time or when they happen in the writing process (e.g., spread of revisions) were also desired. For this, temporal analyses as well as triangulating the data with contextual information is necessary. Lastly, the results showed that the level at which these indicators were discussed as well as the terminology used, differed between the groups of stakeholders. For example, students focused on lower-level behavioral indicators, while teachers focused on higher-level cognitive indicators. This shows that for providing feedback to students, these lower-level behavioral

indicators are still necessary, and additional effort needs to be taken (e.g., additional explanations from teachers) to bring students' understanding of the writing process to these higher levels that are desired by teachers.

8.1.2 Determining capabilities of keystroke analysis

With insight into which indicators are desired, we thereafter determined what is technically feasible with keystroke data, as specified in subquestion 2: *What keystroke features can be used to gain insight into students' writing processes?*

In Chapter 3 we determined the sensitivity of the keystrokes towards differences in writing tasks. Two datasets with both two tasks were compared: academic summary versus copy task and email-writing versus copy task. The results showed that several keystroke features, such as the standard deviation of the interkeystroke interval (IKI) within words, were insensitive to differences between the writing tasks in both datasets. Some features only varied for the academic summary versus copy task; the tasks with a larger difference in complexity or cognitive load. In particular, the mean and standard deviation of the IKI between words, time between keys, and time between subsentences, as well as the number of words only differed between the academic summary task and the copy task, but not between the copy task and the email writing task. Lastly, mean time between words and features related to the task as a whole, such as the number of keystrokes, the number of backspaces, efficiency (characters in the final product per character typed), largest IKI, and total time, were different between the tasks in both datasets. Therefore, these features are already sensitive to small differences in cognitive load or task complexity, and hence might be used to gain insight into students' writing processes.

In Chapter 4 we determined whether the keystroke features commonly extracted in writing research can also be used to predict writing quality, and specifically to predict writing quality early on in the writing process. This could provide insight into which features may be effective in improving writing quality, when providing feedback on the writing process. However, the results showed that the relation between the keystroke features and writing quality was limited. None of the models were able to outperform the baseline in predicting final grade (regression) after the full writing process was finished. In addition, the models for predicting pass/fail (classification) only slightly outperformed the baseline. For these classification models it was shown that the importance of the keystroke features for predicting writing quality differed over time in the writing process. This again stresses

the importance to analyze the keystroke features in relation to time. Moreover, this shows that the relation between keystroke features might be less straight-forward compared to what previous studies found. Therefore, in the remainder of the dissertation I focused on using keystroke logging to model writing *processes* as opposed to writing *quality*.

8.1.3 Gaining insights

For the modeling of the writing process, I specifically focused on revision processes, as the stakeholders in Chapter 2 identified revision as one of the most desired processes to gain insight into, and these processes are rather directly observable in the keystroke data. This was specified as subquestion 3: *How can we model keystroke features to gain insight into students' revision processes?*

In Chapter 5 we provided a product-oriented and process-oriented tagset of revisions. In this tagset, revisions were modeled using ten properties of revisions: processing, trigger, orientation, evaluation, action, linguistic domain, spatial location, temporal location, duration, and sequencing. The results showed how this tagset could be used to annotate a dataset of revisions obtained from keystroke data on these different properties of revision. A majority of the properties were automatically annotated, using rule-based algorithms. Replays of the keystroke data in combination with eye tracking were used to manually annotate the remaining features. To fully automate the analysis of revision processes, these manual features may be learned using machine learning techniques.

Chapter 6 showed that this automated classification is indeed possible for one of the subcategories of orientation in the revision tagset: typos. This allows us to filter or analyze these lower-level type of revisions separately. A process-based model on keystroke data could be trained to distinguish typographic error revisions from other types of revision, especially within a copy task. However, the results did not generalize well to other writing tasks. Hence, for the modeling of keystroke features to gain insight into revision processes, we need to further improve the accuracy of the machine learning models to automatically learn the manual labels within the revision tagset.

8.1.4 Operationalizing insights

Lastly, we got back to the educational stakeholders, to determine how we could use the models obtained in the previous chapters to present information on students' revision pro-

cesses in an interpretable and actionable way. This was specified as subquestion 4: *How can we visualize students' revision processes in order to make them actionable for teachers?*

In Chapter 7, the annotated dataset of revisions was used to create visualizations of students' revision processes. A human-centered design approach with writing researchers and writing teachers was used to transform these visualizations into an interpretable and actionable dashboard on students' revisions. It was shown that especially information on the depth, linguistic domain, temporal location, spatial location, and sequencing of the revisions were considered useful for educational practice. The resulting dashboard proved to be interpretable, with little clarifying questions needed. After interpreting the dashboard, the teachers envisioned a wide variety of actions, ranging from individual coaching to classroom-wide instruction. Hence, the dashboard opens new perspective to bring writing analytics to the classroom, by inviting students to reflect on concrete representations of their writing processes.

8.2 Limitations and future work

The specific limitations for the individual studies can be found in the corresponding chapters. In the following, I describe some overarching limitations and indicate how these may be addressed in future work.

First, this dissertation followed a writing analytics approach. Buckingham Shum et al. (2016) describe the purpose of writing analytics as to better understand writing processes and writing products. In this dissertation, this purpose has been fulfilled by using keystroke logging to gain insight into students' writing processes. Yet, since I posited writing analytics as a subfield of learning analytics, the eventual goal of writing analytics can be considered as to improve learning and teaching of writing (cf. Clow, 2013; Romero & Ventura, 2013). In this light, this dissertation is limited in that I did not actually implement the insights into the writing process in educational practice. As I showed throughout this dissertation, these implementations require many iterations and are hence out of the scope of the current dissertation. However, it remains unknown whether the insights gained from the current dissertation are actually useful for improving learning and teaching of writing. Therefore, a natural follow-up on the research in this dissertation is to implement these insights within educational practices.

The effect of these implementations on the improvement of learning and teaching of writing can be measured using intervention studies (Graham & Harris, 2014). From this

dissertation, for instance, an intervention could be designed based upon the revision dashboard. From Chapter 7, two main types of interventions emerged: class-based instruction and individual coaching sessions. Within a possible class-based instruction intervention, the teacher explains the writing process to increase students' awareness on the different strategies possible, followed by group-based reflection in which students reflect on their writing processes and try to find ways to change their writing process. In individual coaching sessions, the coach would go back to parts in the writing process or product where the student struggled, to identify why the student struggled, and to provide advice or set goals for the next session. As the learning setting will have a large impact on the effectiveness of the intervention, it is important to evaluate the dashboard and the actions upon the dashboard within its educational context, the classroom or individual coaching setting, over a longer period of time (cf. Martinez-Maldonado et al., 2015). Moreover, it is important to clearly describe this context when discussing the results (for a scheme on reporting the content and structure of writing interventions, see Rijlaarsdam et al., 2017).

These interventions can be used to examine whether the *envisioned* actions (Chapter 7) will actually be taken by the teachers, and whether the desirable indicators for providing feedback (Chapter 2) are also *effective* for providing feedback. To measure the effectiveness of the intervention, not only the impact on writing quality needs to be measured, but also the impact on the writing process (e.g., using the revision tagset in Chapter 5). In this way, evidence can be gained on whether students also followed the goals they set and changed their writing process accordingly (cf. Verbert et al., 2014).

Additionally, although the approach in this dissertation takes a cursory account of the temporal aspects of the keystroke data, it warrants more attention. Chapter 6 and previous work showed that writing processes themselves, and the relation between writing processes and writing quality, vary over time (Breetvelt et al., 1994). However, as is common in keystroke analysis, most of the analyses in this dissertation focused on frequencies, while ignoring the sequential nature of the keystroke log. Temporal analysis can provide more detailed insights, (e.g., the identification of frequent temporal patterns), as opposed to the mere counting of events. Accordingly, these temporal aspects are important to take into account when providing automated and personalized insights into students' writing processes. One of the visualizations in the revision dashboard represents revisions in function of time. However, these temporal patterns were not further analyzed.

To date, a paucity of research has examined the temporal aspects of the keystroke data, with some exceptions (Guo et al., 2019; Likens et al., 2017; Xu, 2018; M. Zhang et al., 2016). This could be caused by the fact that the smallest unit in keystroke data, a keystroke, might be too small to be of interest for temporal analysis or pattern mining. For example, the finding that the most common subsequences in writing are sequences of one or more character insertions (Zhu et al., 2019), is, as the authors rightly notice, "intuitive and less interesting". Therefore, more insightful units of analysis need to be identified and extracted first, to allow for meaningful sequential analyses. This can be operationalized, for example, by combining keystroke features into aggregated constructs (cf. Galbraith & Baaijen, 2019). For future work, I envision three potential candidates: bursts (sequences of language bursts without long pause, e.g., P-bursts and R-burst; Baaijen et al., 2012); episodes of linearity (sequences of linear production, without moving elsewhere in the text; Baaijen et al., 2012); and revision events (sequences of keystrokes that constitute a revision, see Chapter 5). For the latter, the current dissertation provides automatic ways of identifying different types of revision events (see Chapter 5). Analyzing the sequential nature of these revision events, using pattern mining techniques, may unravel the recursiveness of revisions, and may be used to determine the effect of specific types of revision (e.g., typos) on the dynamics and (dis)fluency of writing.

Lastly, in this dissertation I did not focus on a specific genre or language background (e.g., L1 versus L2), as I aimed to provide stakeholders with insight into writing processes, without limiting the research to a specific genre or language. Accordingly, this dissertation provides a general way to provide educational stakeholders with insight into writing processes. However, it is unknown how these insights generalize to different contexts. For example, Chapter 7 showed that the interpretation and hence the envisioned actions upon the information in the revision dashboard vary for different tasks (e.g., short summary versus argumentative writing) and for different versions of the task (e.g., draft versus final). Accordingly, future work should identify how these insights into the writing process are used in different contexts. This could be studied, for example, by tailoring the revision tagset to specific genres, languages, and tasks, or to determine whether different contexts result in different actions upon the revision dashboard.

8.3 Reflections and implications of keystroke logging

The primary focus on keystroke logging in my dissertation, in combination with the variety of methods employed, enables me to reflect on the implications of using keystroke logging in educational practices as well as in writing analytics research. Four main themes are described: (1) Reflections on using keystroke data for educational practice, (2) Reflections on relating keystroke data to writing theory, (3) Reflections on keystroke data collection, and (4) Reflections on keystroke data analysis. For each of these reflections, I provide several recommendations for using keystroke data, to leverage the full potential of these data.

8.3.1 Reflections on using keystroke data for educational practice

Throughout this dissertation, we showed that there is no single effective writing process for every writer and every task. In Chapter 4 we found that the relation between the writing process and writing quality is limited. Likewise, Chapter 7 showed that teachers were struggling with qualifying the writing processes; they found it hard to indicate which student showed the 'best' revision process. This means that there is no one-size-fits-all approach for improving students' writing processes. For this, computer-based writing support tools, providing automated and personalized insights into students' writing process can form a scalable solution. These insights can be used by students to increase awareness and reflection on their writing processes (cf. Verbert et al., 2014) and, perhaps with some additional help from teachers or peers, can make students more strategic in their writing (cf. Graham et al., 2020). In Chapter 5 and Chapter 6, we already showed some ways to gain automated and personalized insight into revision processes. Future work should improve the accuracy of automated and personalized insight into these revision processes, and expand beyond revisions, to gain automated insight into multiple aspects of students' writing processes, for instance, planning techniques, source interaction, and session management.

For using keystroke logging in the classroom, the appropriate software needs to be available to collect and (automatically) analyze or visualize the keystroke data. Several tools exist, such as Trace-it, ScriptLog, Inputlog, CyWrite, and EyeWrite (Van Waes et al., 2012). However, most tools make use of low-feature text editors. Although this makes data collection much more consistent and accurate, and hence preferable for writing research, this is not optimal for teaching writing, as these are not the tools students are used to. One exception is Inputlog (Leijten & Van Waes, 2013), which works as a plug-in to Microsoft Word

(at the cost of some minimal loss in accuracy of the data collection). However, Inputlog is only available for Windows computers, which restricts students in their freedom of operating system use. This does not necessarily have to prevent the use of keystroke logging in the classroom. Especially for shorter assignments, which still can be used to provide students with insight into their writing process, computer classrooms could be booked to make sure everyone has access to the tool. For longer-term assignments, such as writing of a thesis, I propose to develop a keystroke logging tool that could be used as a plug-in to the software the student is familiar with, on the desktop or laptop they are familiar with. For example, this could be a plug-in for Google Docs, which could be used with any operating system and has significant overlap with word-processing tools students are familiar with.

In addition, the analyses provided by these tools are often aimed at researchers with high data literacy, and hence are less accessible for teachers (and students). This dissertation demonstrated that human-centered design of these analytics tools is one solution to make these tools more interpretable and actionable for educational stakeholders. For example, as shown in Chapter 7, teachers are concerned with the multiplicity of the tools they need to operate. Therefore, future work should consider possibilities to integrate analyses and visualizations of the writing process within existing tools, such as the university's learning management system.

My last reflection on using keystroke logging in the educational contexts is related to ethical concerns. Keystroke data are considered biometric data and hence could be used to identify and authenticate students (Karnan et al., 2011). On the one hand, these biometric data can be highly useful in detecting plagiarism, or detecting whether it was really student X who wrote that assignment or online exam (and not their roommate). This advantage has already been used in massive open online courses (e.g., `http://coursera.org`) or within universities (e.g., `http://cadmus.io`). On the other hand, these biometric data also have the risk of breaching privacy, especially if combined with other data. Accordingly, for using the data for educational purposes, the data need to be safely stored. In addition, educational stakeholders and researchers need to make sure to not collect sensitive information within the keystroke data (e.g., Facebook passwords) or collect data when students do not want to be tracked. At the very least, it needs to be possible to remove those data. Overall, we need to be transparent about what data is collected, when it is collected, and how it used. Several frameworks to do so can be borrowed from the learning analytics

community (see e.g., Greller & Drachsler, 2012; Slade & Prinsloo, 2013) or the fairness, accuracy, and transparency (FAT) community in machine learning and artificial intelligence (see e.g., `http://fatml.org`). Given the biometric nature of keystroke data, I suggest for future work to determine how these frameworks can be applied to using keystroke logging in educational contexts.

8.3.2 Reflections on relating keystroke data to writing theory

In this dissertation, I have used a data-driven approach, combined with theory on writing processes and insights from educational stakeholders. Accordingly, I did not aim to make any theoretical claims. However, I do feel that such a data-driven approach, informed by theory on writing processes and insights from educational stakeholders, may provide insight into theories of writing, in three ways.

First, as shown in Chapter 5, data-driven approaches can be used to automatically identify different categories in the revision tagset which are considered to be of theoretical interest, without time-intensive manual annotations needed. Likewise, these approaches could be used to automatically remove certain categories from the analysis, which might be of less interest, or which might confound the analysis, such as typographic error revisions (see Chapter 6). This can result in more robust analyses of only the specific constructs of interest.

Second, data-driven approaches could, on a large scale, identify points of interest in the keystroke log, such as points where many revisions are made. Alternatively, given the larger sample sizes, it may be possible to find more subtle differences between groups of writers. In addition, a high predictive power of a keystroke feature, found in multiple studies or contexts, might indicate a relationship worth investigating. These points of interest may in turn be used as a basis for theory-driven empirical studies, for example, by triangulating the data with manual annotations, eye tracking, and/or thinking-aloud, to further explore the causes and the theoretical underpinning of these points of interest.

Lastly, data-driven approaches could shed light on constructs that are hard to theoretically define. For example, the constructs of non-linearity or fluency in writing have many different definitions (Van Waes & Leijten, 2015). A data-driven approach might provide more insight into what processes are seen as 'non-linear' or 'disfluent'. In addition, this could even be used to rank and or cluster writing processes on their fluency, or non-

linearity, for instance, in the context of more complex, multi-session text composition (e.g., novel writing).

8.3.3 Reflections on keystroke data collection

There are ample possibilities to log keystroke data. However, when using keystroke logging in writing research, several aspects of the data collection need to be considered. First of all, the accuracy of the data plays a role. The accuracy of the hardware used (e.g., keyboard polling rate) influences the preciseness of the claims that can be made. In addition, the data collected across tools are not always in similar formats or similar granularity. For example, mouse data are not always collected. Mouse data are very important to determine the location of the keystroke within the written product, for example, when writers go back in the text (with their mouse) to add some characters. They are crucial when accurately determining what has been deleted, for example, when someone selects a whole sentence, and deletes this (which only results in one delete key).

Second, some additional information may need to be collected, depending on the research question and research design. As keystrokes are very sensitive to other factors, it is advisable to account for individual differences, for example by including information on the device used, the keyboard layout, the handedness, and the typing style as control variable. To control for typing skills, a copy task could be used as a baseline (Van Waes et al., 2019). Alternatively, multi-level analysis could be used to account for the hierarchical nature of the data. In addition, other types of information could be added to keystroke data, to get a better understanding of why certain behaviors are found (e.g., long pauses). This could be done, for example, by logging source usage (Leijten, Van Waes, et al., 2019) or using eye tracking to identify what a writer is looking at within a long pause or before a revision (Chukharev-Hudilainen et al., 2019).

Third, several keystroke logging tools transform the raw keystroke data into so-called keystroke features, such as IKI or revision burst. However, similar features can be implemented differently by different tools or researchers (see Chapter 4). In addition, different thresholds are used, for example, for pause timings often a threshold is used of 2000 ms (Kaufer et al., 1986; Wengelin, 2006), but other thresholds are used as well, such as 500 ms (Chukharev-Hudilainen, 2014) or the mean IKI plus two times the *SD* of the IKI (Deane, 2014). These different definitions can result in different outcomes, and make it hard to compare the effects of the features across studies (Xu & Ding, 2014). Accordingly, future

work should be more transparent in which operationalization of the keystroke feature was used (cf. Chapter 4). Moreover, it is important to acknowledge that the keystroke logging tool or a fixed threshold specified in previous work might not always provide the best operationalization of a specific keystroke feature for the research question at hand.

Fourth, to further the use of keystroke analysis, more research data need to be openly available. There are already some open keystroke datasets available (e.g., Monaco et al., 2012; Tappert et al., 2010; Dhakal et al., 2018), but this is still very limited. Sharing these sensitive keystroke data requires that the data are anonymized, for example by reporting aggregated measures. In addition, to increase the accessibility of these data, it would be advisable to further develop a standardized format for these log files (cf. Van Horenbeeck et al., 2015).

Lastly, current data collections tend to be uniformal: they usually only consider relatively short writing tasks that are written within a single session, using a keyboard. However, writing is generally spread out over multiple sessions and can include multiple modes, such as handwriting and typing (Leijten et al., 2014). In addition, writing is omnipresent (Brandt, 2014) and written on (and retrieved via) different devices, resulting in different input modes, such as smartphone keyboards and touch interfaces. Accordingly, future work should also collect data in these settings, to reflect writers' multimodal text production. In this way, we can identify how information flows between different modes of text production, and how this influences our overall writing process.

8.3.4 Reflections on keystroke data analysis

To make (valid) inferences about the fine-grained keystroke data, other types of analyses may be preferred, compared to the traditional statistical analyses that are common in the field of writing research. Here I provide three recommendations for analyzing keystroke data.

First, it is important to identify the features of interest, before analyzing the data. Several keystroke logging tools provide large lists of keystroke features that are calculated based on the raw keystroke data. Given this large list of features, it is highly likely that statistical analyses will show significant effects in at least some of these features. However, these are likely to include false positives (i.e., type II errors). In addition, these features will probably be non-independent, violating the assumption of non-multicollinearity made by many statistical analyses. Therefore, it is important to only include features that are useful to

your research question at hand or to choose for machine learning algorithms, where the outcome of the prediction is more important than inferences about the features included.

Second, feature reduction or the combination of keystroke features might result in findings that are more robust or easier to relate to the writing process. Individual keystrokes features have been shown very sensitive to small differences in tasks (Chapter 3). In addition, individual keystroke features can be related to multiple underlying cognitive processes. For example, pauses could indicate sentence planning, global text planning, and lexical access, which to some extent can be related to the different locations of the pauses (Medimorec & Risko, 2017). A backspace key could indicate a typo or slip of the finger, or a thoughtful revision of meaning (Chapter 6). Consequently, in isolation, it is hard to relate a single keystroke feature to specific cognitive processes, or to relate a single feature to constructs such as task complexity or writing quality (Galbraith & Baaijen, 2019). Therefore, as Galbraith & Baaijen (2019) argue, it is valuable to aggregate and/or combine keystroke features (e.g., using factor analysis or dimensionality reduction), as these aggregated (latent) variables could capture more general properties of the writing process.

Lastly, keystroke data are very noisy and usually not normally distributed. This requires specific types of machine learning algorithms, or more advanced statistical methodology. Typically, analyses are done on means and medians of pause timings such as IKI. In this way, much information is lost, which might result in higher bias in the model estimates (Roeser et al., 2020). Preferred approaches include: (1) analyzing IKIs at different text boundaries (e.g., between words, sentences; Galbraith & Baaijen, 2019; Medimorec & Risko, 2017); (2) allowing IKIs to stem from multiple distributions, e.g., using mixture models (Galbraith & Baaijen, 2019; Roeser et al., 2019); (3) modeling character bigrams as random intercept in multi-level analysis, to account for bigram variance (Roeser et al., 2020); and (4) using autoregression models, in which each IKI is also regressed based on the preceding IKI (Roeser et al., 2020). In these ways, more justice is done to the nature of these data, which can result in more accurate findings.

8.4 Conclusion

To conclude, this dissertation expands our knowledge on the use of keystroke logging in writing research. Specifically, it answered the main research question:

How can keystroke logging be used to gain meaningful insight into students' writing processes?

This dissertation showed that, to gain meaningful insight into students' writing processes, it is important to consider the technical possibilities of the keystroke data. Using data-driven approaches, automatic and scalable insights may be gained into students' revision processes. In addition, by using a human-centered approach, these insights may be transformed into *meaningful* insights, in the form of a revision dashboard. With this, I illuminated some of the Keys to Writing. ■

References

Abdel Latif, M. M. (2009). Toward a new process-based indicator for measuring writing fluency: Evidence from L2 writers' think-aloud protocols. *Canadian Modern Language Review*, *65*(4), 531–558. https://doi.org/10.3138/cmlr.65.4.531

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ, US: Wiley. https://doi.org/10.1007/978-3-642-04898-2_161

Alamargot, D., & Chanquoy, L. (Eds.). (2001). *Through the models of writing (Studies in Writing)* (Vol. 9). Dordrecht, The Netherlands: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-010-0804-4

Alamargot, D., Dansac, C., Chesnet, D., & Fayol, M. (2007). Parallel processing before and after pauses: A combined analysis of graphomotor and eye movements during procedural text production. In M. Torrance, L. Van Waes, & D. Galbraith (Eds.), *Writing and cognition: Research and applications* (Vol. 20, pp. 13–29). Leiden, The Netherlands: Brill. https://doi.org/10.1163/9781849508223_003

Allal, L. (2000). Metacognitive regulation of writing in the classroom. In A. Camps & M. Milian (Eds.), *Metalinguistic activity in learning to write* (Vol. 6, pp. 145–166). Amsterdam, The Netherlands: Amsterdam University Press.

Allal, L., Chanquoy, L., & Largy, P. (Eds.). (2004). *Revision cognitive and instructional processes (Studies in Writing)* (Vol. 13). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-94-007-1048-1

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. In *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 22–29). Retrieved from https://eric.ed.gov/?id=ED592674

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Computer-based writing instruction. *Handbook of writing research*, 316–329. Retrieved from https://eric.ed.gov/?id=ED586512

Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S., & McNamara, D. S. (2016). Investigating boredom and engagement during writing using multiple sources of information: the

essay, the writer, and keystrokes. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 114–123). https://doi.org/10.1145/2883851.2883939

Alves, R. A., Castro, S. L., De Sousa, L., & Strömqvist, S. (2007). Influence of typing skill on pause-execution cycles in written composition. In M. Torrance, L. Van Waes, & D. Gabraith (Eds.), *Writing and cognition: Research and applications* (Vol. 20, pp. 55–65). Leiden, The Netherlands: Brill. https://doi.org/10.1163/9781849508223_005

Alves, R. A., Castro, S. L., & Olive, T. (2008). Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International Journal of Psychology*, *43*(6), 969–979. https://doi.org/10.1080/00207590701398951

Alves, R. A., & Limpo, T. (2015). Progress in written language bursts, pauses, transcription, and written composition across schooling. *Scientific Studies of Reading*, *19*(5), 374–391. https://doi.org/10.1080/10888438.2015.1059838

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3). https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Baaijen, V. M., & Galbraith, D. (2018). Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction*, *36*(3), 1–25. 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.*

Banerjee, R., Feng, S., Kang, J. S., & Choi, Y. (2014). Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1469–1473). https://doi.org/10.3115/v1/D14-1155

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. https://doi.org/10.1111/modl.12316

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. https://doi.org/10.1016/j.jml.2007.09.002

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint*. Retrieved from https://arxiv.org/abs/1506.04967

Beauvais, C., Olive, T., & Passerault, J.-M. (2011). Why are some texts good and others not? Relationship between text quality and management of the writing processes. *Journal of Educational Psychology*, *103*(2), 415–428. https://doi.org/10.1037/a0022545

Becker, A. (2006). A review of writing model research based on cognitive processes. In A. Horning & A. Becker (Eds.), *Revision: History, theory, and practice* (pp. 25–49). Retrieved from https://wac.colostate.edu/books/horning_revision/chapter3.pdf

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with Python*. O'Reilly Media Inc.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, *14*(3), 191–205. https://doi.org/10.1016/j.jslw.2005.08.001

Bixler, R., & D'Mello, S. (2013). Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI)* (pp. 225–234). New York, NY, US: ACM. https://doi.org/10.1145/2449396.2449426

Bowen, N., & Van Waes, L. (2020). Exploring revisions in academic text: Closing the gap between process and product approaches in digital writing. *Written Communication*, *37*(3). https://doi.org/10.13140/RG.2.2.15312.46086

Boytsov, L. (2011). Indexing methods for approximate dictionary searching: Comparative analysis. *Journal of Experimental Algorithmics*, *16*, 1–91. https://doi.org/10.1145/1963190.1963191

Braaksma, M. A., Rijlaarsdam, G., Van den Bergh, H., & Van Hout-Wolters, B. H. M. (2004). Observational learning and its effects on the orchestration of writing processes. *Cognition and Instruction*, *22*(1), 1–36. https://doi.org/10.1207/s1532690Xci2201_1

Brandt, D. (2014). *The rise of writing: Redefining mass literacy*. Cambridge, UK: Cambridge University Press.

Breetvelt, I., Van den Bergh, H., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction*, *12*(2), 103–123. Retrieved from https://www.jstor.org/stable/3233677

Buckingham Shum, S., Ferguson, R., & Martinez-Maldonado, R. (2019). Human-centred learning analytics. *Journal of Learning Analytics*, *6*(2), 1–9. https://doi.org/10.18608/jla.2019.62.1

Buckingham Shum, S., Knight, S., McNamara, D., Allen, L., Bektik, D., & Crossley, S. (2016). Critical perspectives on writing analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 481–483). ACM. https://doi.org/10.1145/2883851.2883854

Casey, K. (2017). Using keystroke analytics to improve pass-fail classifiers. *Journal of Learning Analytics*, *4*(2), 189–211. https://doi.org/10.18608/jla.2017.42.14

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, *12*(3), 267–296. https://doi.org/10.1016/S1060-3743(03)00038-9

Chanquoy, L. (1997). Thinking skills and composing: Examples of text revision. In J. H. M. Hamers & M. Overtoom (Eds.), *Inventory of european programmes for teaching thinking* (pp. 179–185). Utrecht, The Netherlands: Sardes.

Charleer, S., Klerkx, J., Duval, E., De Laet, T., & Verbert, K. (2016). Creating effective learning analytics dashboards: Lessons learnt. In *European conference on technology enhanced learning* (pp. 42–56). https://doi.org/10.1007/978-3-319-45153-4_4

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

Choi, Y. H. (2007). On-line revision behaviors in EFL writing process. *English Teaching*, *62*(4), 69–93.

Chukharev-Hudilainen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research*, *6*(1), 61–84. https://doi.org/10.17239/jowr-2014.06.01.3

Chukharev-Hudilainen, E. (2019). Empowering automated writing evaluation with keystroke logging. In E. Lindgren & K. Sullivan (Eds.), *Observing writing* (Vol. 38, pp. 125–142). Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004392526_007

Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., & Feng, H.-H. (2019). Combined deployable keystroke logging and eyetracking for investigating L2 writing fluency. *Studies in Second Language Acquisition*, *41*(3), 583–604. https://doi.org/10.1017/S027226311900007X

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683–695. https://doi.org/10.1080/13562517.2013.827653

Conde, M. A., & Hernández-García, A. (2015). Learning analytics for educational decision making. *Computers in Human Behavior*, *47*, 1–3. https://doi.org/10.1016/j.chb.2014.12.034

Conijn, R., Cook, C., van Zaanen, M., & Van Waes, L. (under review). *Early prediction of writing quality using keystroke logging.*

Conijn, R., Dux Speltz, E., van Zaanen, M., Van Waes, L., & Chukharev-Hudilainen, E. (2020). A process-oriented dataset of revisions during writing. In *Proceedings of the 12th language resources and evaluation conference* (pp. 356–361). Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.44

Conijn, R., Dux Speltz, E., van Zaanen, M., Van Waes, L., & Chukharev-Hudilainen, E. (under review). *A product and process oriented tagset for revisions in writing.*

Conijn, R., Martinez-Maldonado, R., Knight, S., Buckingham Shum, S., Van Waes, L., & van Zaanen, M. (under review). *How to provide automatic feedback on the writing process? A participatory approach to writing analytics design.*

Conijn, R., Nij Bijvank, W., Snijders, C., Kleingeld, A., & Matzat, U. (2018). From raw to ready-made data. a hands-on manual for pre-processing Learning Management System log data for learning analytics. In C. Stuetzer, M. Welker, & M. Egger (Eds.), *Computational Social Science in the Age of Big Data: Concepts, Methodologies, Tools, and Applications.* Cologne, Germany: Herbert von Halem Verlag.

Conijn, R., Roeser, J., & van Zaanen, M. (2019). Understanding the keystroke log: The effect of writing task on keystroke features. *Reading and Writing*, *32*(9), 2353–2374. https://doi.org/10.1007/s11145-019-09953-8

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Conijn, R., Van den Beemt, A., & Cuijpers, P. (2018). Predicting student performance in a blended MOOC. *Journal of Computer Assisted Learning*, *34*(5), 615–628. https://doi.org/10.1111/jcal.12270

Conijn, R., Van der Loo, J., & van Zaanen, M. (2018). What's (not) in a keystroke? Automatic discovery of students' writing processes using keystroke logging. In *Companion Proceedings of the 8th International Conference on Learning Analytics & Knowledge.* Sydney, Australia.

Conijn, R., Van Waes, L., & van Zaanen, M. (2020). Human-centered design of a dashboard on students' revisions during writing. In *Conference proceedings of the 14th European Conference on Technology Enhanced Learning, EC-TEL* (pp. 1–15). https://doi.org/10.1007/978-3-030-57717-9_3

Conijn, R., & van Zaanen, M. (2017a). Identifying writing tasks using sequences of keystrokes. In *Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning* (pp. 28–35). Eindhoven, The Netherlands.

Conijn, R., & van Zaanen, M. (2017b). Trends in student behavior in online courses. In *Proceedings of the 3rd International Conference on Higher Education Advances* (pp. 649–657). Valencia, Spain. https://doi.org/10.4995/HEAd17.2017.5337

Conijn, R., van Zaanen, M., Leijten, M., & Van Waes, L. (2019). How to typo? Building a process-based model of typographic error revisions. *Journal of Writing Analytics*, *3*, 69–95.

Conijn, R., van Zaanen, M., & Van Waes, L. (2019). Don't wait until it is too late: The effect of timing of automated feedback on revision in esl writing. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Learning with Meaningful Technologies, Conference proceedings of the 14th European Conference on Technology Enhanced Learning, EC-TEL* (pp. 577–581). Delft, The Netherlands. https://doi.org/10.1007/978-3-030-29736-7_43

Cook, C., Conijn, R., Antheunis, M., & Schaafsma, J. (2019). For whom the gamer trolls: empirical model of trolling in the online gaming context. *Journal of Computer-Mediated Communication*. https://doi.org/10.1093/jcmc/zmz014

Cotos, E. (2015). Automated writing analysis for writing pedagogy: From healthy tension to tangible prospects. *Writing and Pedagogy*, *6*, 1–29. https://doi.org/10.1558/wap.v7i2-3.26381

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

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), 415-443. https://doi.org/10.17239/jowr-2020.11.03.01

Crossley, S., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14–27. https://doi.org/10.3758/s13428-018-1142-4

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 Psy-*

chology: Learning, Memory, and Cognition, *36*(6), 1369–1380. https://doi.org/10.1037/a0020696

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

Daxenberger, J., & Gurevych, I. (2012). A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proceedings of COLING 2020: Technical papers* (pp. 711–726).

Daxenberger, J., & Gurevych, I. (2013). Automatically classifying edit categories in Wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 578–589).

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Deane, P. (2014). Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks. *ETS Research Report Series*(1), 1–23. https://doi.org/10.1002/ets2.12002

Dhakal, V., Feit, A. M., Kristensson, P. O., & Oulasvirta, A. (2018). Observations on typing from 136 million keystrokes. In *Conference on Human Factors in Computing Systems-Proceedings* (pp. 1–12). https://doi.org/10.1145/3173574.3174220

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1), 1–35. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1640

Dollinger, M., Liu, D., Arthars, N., & Lodge, J. (2019). Working together in learning analytics towards the co-creation of value. *Journal of Learning Analytics*, *6*(2), 10–26. https://doi.org/10.18608/jla.2019.62.2

Donnelly, S., & Verkuilen, J. (2017). Empirical logit analysis is not logistic regression. *Journal of Memory and Language*, *94*, 28–42. https://doi.org/10.1016/j.jml.2016.10.005

Drost, H.-G. (2018). philentropy: Similarity and distance quantification between probability functions [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=philentropy (R package version 0.3.0)

Dvorak, A., Merrick, N. L., Dealey, W. L., & Ford, G. C. (1936). *Typewriting behavior*. Oxford, England: American Book Co.

El Ebyary, K., & Windeatt, S. (2010). The impact of computer-based feedback on students' written work. *International Journal of English Studies*, *10*(2), 121–142. https://doi.org/10.6018/ijes/2010/2/119231

Emerson, A., Smith, A., Smith, C., Rodríguez, F. J., Min, W., Wiebe, E. N., ... Lester, J. C. (2019). Predicting early and often: Predictive student modeling for block-based programming environments. In *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 39–48).

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

Ferguson, P. (2011). Student perceptions of quality feedback in teacher education. *Assessment & Evaluation in Higher Education*, *36*(1), 51–62. https://doi.org/10.1080/02602930903197883

Ferris, D. (2011). Responding to student errors: Issues and strategies. In D. Ferris (Ed.), *Treatment of error in second language student writing* (pp. 49–76). Ann Arbor, MI, US: University of Michigan Press.

Fidalgo, R., & Torrance, M. (2017). Developing writing skills through cognitive self-regulation instruction. In R. Fidalgo, K. Harris, & M. Braaksma (Eds.), *Design principles for teaching effective writing* (Vol. 34, pp. 89–118). Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004270480_006

Figueredo, L., & Varnhagen, C. K. (2006). Spelling and grammar checkers: are they intrusive? *British Journal of Educational Technology*, *37*(5), 721–732. https://doi.org/10.1111/j.1467-8535.2006.00562.x

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research*, *57*(4), 481–506. https://doi.org/10.3102/00346543057004481

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

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387. https://doi.org/10.2307/356600

Flower, L., Hayes, J. R., Carey, L., Schriver, K., & Stratman, J. (1986). Detection, diagnosis, and the strategies of revision. *College Composition and Communication*, *37*(1), 16–55. https://doi.org/10.2307/357381

Gabriska, D., & Ölveckỳ, M. (2018). Issues of adaptive interfaces and their use in educational systems. In *16th International Conference on Emerging eLearning Technologies and Applications (ICETA)* (pp. 173–178). IEEE. https://doi.org/10.1109/ICETA.2018.8572096

Gabry, J., & Goodrich, B. (2016). rstanarm: Bayesian applied regression modeling via Stan [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rstanarm (R package version 2.13.1)

Galbraith, D., & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In E. Lindgren & K. Sullivan (Eds.), *Observing writing* (Vol. 38, pp. 306–325). Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004392526_015

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL, US: CRC Press.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Giacomin, J. (2014). What is human centred design? *The Design Journal*, *17*(4), 606–623. https://doi.org/10.2752/175630614X14056185480186

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, *20*(4), 304–315. https://doi.org/10.1016/j.learninstruc.2009.08.007

Giot, R., El-Abed, M., & Rosenberger, C. (2009). Keystroke dynamics authentication for collaborative systems. In *International Symposium on Collaborative Technologies and Systems* (pp. 172–179). https://doi.org/10.1109/CTS.2009.5067478

Grabowski, J. (2008). The internal structure of university students' keyboard skills. *Journal of Writing Research*, *1*(1), 27–52. https://doi.org/10.17239/jowr-2008.01.01.2

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

Graham, S., Bañales, G., Ahumada, S., Muñoz, P., Alvarez, P., & Harris, K. R. (2020). Writing strategies interventions. In D. L. Dinsmore, L. K. Fryer, & M. M. Parkinson (Eds.), *Handbook of strategies and strategic processing* (pp. 141–158). London, UK: Routledge.

Graham, S., & Harris, K. R. (2014). Conducting high quality writing intervention research: Twelve recommendations. *Journal of Writing Research*, *6*(2), 89–123. https://doi.org/10.17239/jowr-2014.06.02.1

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, *115*(4), 523–547. https://doi.org/10.1086/681947

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, *104*(4), 879–896. https://doi.org/10.1037/a0029185

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*, *15*(3), 42–57. Retrieved from https://www.jstor.org/stable/jeductechsoci.15.3.42

Gunetti, D., & Picardi, C. (2005). Keystroke analysis of free text. *ACM Transactions on Information and System Security (TISSEC)*, *8*(3), 312–347. https://doi.org/10.1145/1085126.1085129

Guo, H., Deane, P. D., Van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement*, *55*(2), 194–216. https://doi.org/10.1111/jedm.12172

Guo, H., Zhang, M., Deane, P., & Bennett, R. E. (2019). Writing process differences in subgroups reflected in keystroke logs. *Journal of Educational and Behavioral Statistics*, *44*(5), 571–596. https://doi.org/10.3102/1076998619856590

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. https://doi.org/10.1207/s15327051hci0402_3

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hausser, J., & Strimmer, K. (2014). entropy: Estimation of entropy, mutual information and related quantities [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=entropy (R package version 1.2.1)

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ, US: Lawrence Erlbaum Associates, Inc.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*(3), 369–388. https://doi.org/10.1177/0741088312451260

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

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

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–ai complementarity. *Journal of Learning Analytics*, *6*(2), 27–52. https://doi.org/10.18608/jla.2019.62.3

Hounsell, D. (1984). Essay planning and essay writing. *Higher Education Research and Development*, *3*(1), 13–31. https://doi.org/10.1080/0729436840030102

Hounsell, D. (1997). Contrasting conceptions of essay-writing. In R. Marton, D. Hounsell, & N. Entwistle (Eds.), *The experience of learning: Implications for teaching and studying in higher education* (pp. 106–125). Edinburgh, UK: Scottish Academic Press.

Itua, I., Coffey, M., Merryweather, D., Norton, L., & Foxcroft, A. (2014). Exploring barriers and solutions to academic writing: Perspectives from students, higher education and further education tutors. *Journal of further and Higher Education*, *38*(3), 305–326. https://doi.org/10.1080/0309877X.2012.726966

Jansen, R. S., Van Leeuwen, A., Janssen, J., Conijn, R., & Kester, L. (2020). Supporting learners' self-regulated learning in massive open online courses. *Computers & Education*, *146*. https://doi.org/10.1016/j.compedu.2019.103771

Janssen, D., Van Waes, L., & Van den Bergh, H. (1996). Effects of thinking aloud on writing processes. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, individual differences, and applications* (pp. 233–250). Mahwah, NJ, US: Lawrence Erlbaum Associates, Inc.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122

Kano, A., Read, J. C., Dix, A., & MacKenzie, I. S. (2007). ExpECT: An expanded error categorisation method for text input. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it* (Vol. 1, pp. 147–156). British Computer Society. https://doi.org/10.1145/1531294.1531314

Karnan, M., Akila, M., & Krishnaraj, N. (2011). Biometric personal authentication using keystroke dynamics: A review. *Applied Soft Computing*, *11*(2), 1565–1573. https://doi.org/10.1016/j.asoc.2010.08.003

Kaufer, D. S., Hayes, J. R., & Flower, L. (1986). Composing written sentences. *Research in the Teaching of English*, *20*(2), 121–140. Retrieved from https://www.jstor.org/stable/40171073

Kellogg, R. T. (1987). Writing performance: Effects of cognitive strategies. *Written Communication*, *4*(3), 269–298. https://doi.org/10.1177/0741088387004003003

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–71). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1–26. https://doi.org/10.17239/jowr-2008.01.01.1

Kennedy, G. E., & Judd, T. S. (2007). Expectations and reality: Evaluating patterns of learning behaviour using audit trails. *Computers & Education*, *49*(3), 840–855. https://doi.org/10.1016/j.compedu.2005.11.023

Kernighan, M. D., Church, K. W., & Gale, W. A. (1990). A spelling correction program based on a noisy channel model. In *Proceedings of the 13th conference on Computational linguistics* (Vol. 2, pp. 205–210). Association for Computational Linguistics. https://doi.org/10.3115/997939.997975

Kidd, P. S., & Parshall, M. B. (2000). Getting the focus and the group: Enhancing analytical rigor in focus group research. *Qualitative health research*, *10*(3), 293–308. https://doi.org/10.1177/104973200129118453

Kim, H.-C. (1996). *Diagnosing and filtering typing error corrections in text revision history* (Tech. Rep. No. TRITA-NA-P9633 IPLab-113). Department of Numerical Analysis and Computing Science, Royal Institute of Technology,Stockholm.

KNAW, NFU, NWO, TO2-federatie, Vereniging Hogescholen, & VSNU. (2018). *The Netherlands Code of Conduct for Research Integrity*. DANS. https://doi.org/10.17026/dans-2cj-nvwu

Knight, S., Abel, S., Shibani, A., Goh, Y. K., Conijn, R., Gibson, A., ... Buckingham Shum, S. (in press). Are you being rhetorical? An open corpus of machine annotated rhetorical moves. *Journal of Learning Analytics*.

Knight, S., Allen, L., Gibson, A., McNamara, D., & Buckingham Shum, S. (2017). Writing analytics literacy: Bridging from research to practice. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 496–497). New York, NY, US: Association for Computing Machinery. https://doi.org/10.1145/3027385.3029425

Kollberg, P. (1996). S-notation as a tool for analysing the episodic structure of revisions. In *European Writing Conferences* (pp. 1–15).

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability. *Departmental Papers (ASC)*, *43*, 1–10. Retrieved from https://repository.upenn.edu/asc_papers/43

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). London, UK: Elsevier.

Kuhn, M. (2019). caret: Classification and Regression Training [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=caret (R package version 6.0–80)

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education*, *23*(2), 157–172. https://doi.org/10.1080/03075079812331380364

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* (Vol. 38, pp. 71–95). Leiden, The Netherlands: Brill. 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

Leijten, M., Van Waes, L., Schriver, K., & Hayes, J. R. (2014). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research*, *5*(3), 285–337. https://doi.org/10.17239/jowr-2014.05.03.3

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

Levy, C. M., & Ransdell, S. (1996). Writing signatures. In *The science of writing: Theories, methods, individual differences and applications* (pp. 149–161). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Levy, M., & Moore, P. (2018). Qualitative research in CALL. *Language Learning & Technology*, *22*(2), 1–7. https://doi.org/10125/44638

Likens, A. D., Allen, L. K., & McNamara, D. S. (2017). Keystroke Dynamics Predict Essay Quality. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci 2017)* (pp. 2573–2578). London, UK.

Lindgren, E., & Sullivan, K. (Eds.). (2019). *Observing writing: Insights from keystroke logging and handwriting (Studies in Writing)* (Vol. 38). Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004392526

Lindgren, E., & Sullivan, K. P. (2006a). Analysing online revision. In K. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing: Methods and applications* (pp. 157–188). Oxford, UK: Elsevier.

Lindgren, E., & Sullivan, K. P. (2006b). Writing and the analysis of revision: An overview. In K. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing: Methods and applications* (pp. 31–40). Oxford, UK: Elsevier.

Lindgren, E., Westum, A., Outakoski, H., & Sullivan, K. P. H. (2019). Revising at the leading edge: Shaping ideas or clearing up noise. In E. Lindgren & K. Sullivan (Eds.), *Observing writing* (Vol. 38, pp. 346–365). Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004392526_017

Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2014). Towards best ESL practices for implementing automated writing evaluation. *Calico Journal*, *31*(3), 323–344. Retrieved from https://www.jstor.org/stable/calicojournal.31.3.323

Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: Aligning learning analytics with learning design. *American Behavioral Scientist*, *57*(10), 1439–1459. https://doi.org/10.1177/0002764213479367

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901–918. https://doi.org/10.1037/a0037123

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Mao, Y., Zhi, R., Khoshnevisan, F., Price, T., Barnes, T., & Chi, M. (2019). One minute is enough: Early prediction of student success and event-level difficulty during novice programming tasks. In *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 119–128).

Martinez-Maldonado, R., Echeverria, V., Fernandez Nieto, G., & Buckingham Shum, S. (2020). From data to insights: A layered storytelling approach for multimodal learning analytics. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–15). https://doi.org/10.1145/3313831.3376148

Martinez-Maldonado, R., Kay, J., Yacef, K., Edbauer, M.-T., & Dimitriadis, Y. (2013). MTClassroom and MTDashboard: Supporting analysis of teacher attention in an orchestrated multi-tabletop classroom. In *Proceedings of the International Conference on Computer-Supported Collaborative Learning (CSCL'13)* (pp. 320–327).

Martinez-Maldonado, R., Pardo, A., Mirriahi, N., Yacef, K., Kay, J., & Clayphan, A. (2015). LATUX: An iterative workflow for designing, validating and deploying learning analytics visualisations. *Journal of Learning Analytics*, *2*(3), 9–39. https://doi.org/10.18608/jla.2015.23.3

Mateos, M., & Solé, I. (2009). Synthesising information from various texts: A study of procedures and products at different educational levels. *European Journal of Psychology of Education*, *24*(4), 435–451. https://doi.org/10.1007/BF03178760

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and Stan*. Boca Raton, FL, US: CRC Press.

Medimorec, S., & Risko, E. F. (2017). Pauses in written composition: On the importance of where writers pause. *Reading and Writing*, *30*(6), 1267–1285. https://doi.org/10.1007/s11145-017-9723-7

Min, H.-T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing*, *15*(2), 118–141. https://doi.org/10.1016/j.jslw.2006.01.003

Monaco, J. V., Bakelman, N., Cha, S.-H., & Tappert, C. C. (2012). Developing a keystroke biometric system for continual authentication of computer users. In *European intelligence and Security Informatics Conference (EISIC)* (pp. 210–216). https://doi.org/10.1109/EISIC.2012.58

Monahan, B. D. (1984). Revision strategies of basic and competent writers as they write for different audiences. *Research in the Teaching of English*, *18*(3), 288–304. Retrieved from https://www.jstor.org/stable/40171020

Murray, D. M. (1978). Internal revision: A process of discovery. In C. Cooper & R. Odell (Eds.), *Research on composing: Points of departure* (pp. 85–103). Urbana, IL, US: National Council of Teachers of English.

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. https://doi.org/10.1111/0026-7902.00007

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, *10*(11), 591–613. https://doi.org/10.1111/lnc3.12207

Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1–15. https://doi.org/10.1037/0033-295X.88.1.1

Norton, L. S. (1990). Essay-writing: what really counts? *Higher Education*, *20*(4), 411–442. https://doi.org/10.1007/BF00136221

NVivo. (2015). *NVivo Qualitative Data Analysis Software.* QSR International Pty Ltd. (Software)

Olive, T. (2014). Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of Writing Research*, *6*(2), 173–194. https://doi.org/10.17239/jowr-2014.06.02.4

Olive, T., & Kellogg, R. T. (2002). Concurrent activation of high-and low-level production processes in written composition. *Memory & Cognition*, *30*(4), 594–600. https://doi.org/10.3758/BF03194960

Olson, J. S., Wang, D., Olson, G. M., & Zhang, J. (2017). How people write together now: Beginning the investigation with advanced undergraduates in a project course. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *24*(1), 1–40. https://doi.org/10.1145/3038919

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing*, *15*(2), 68–85. https://doi.org/10.1016/j.asw.2010.05.004

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

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–587. https://doi.org/10.1177/0265532209340192

Polyzou, A., & Karypis, G. (2019). Feature extraction for next-term prediction of poor student performance. *IEEE Transactions on Learning Technologies*, *12*(2), 237–248. https://doi.org/10.1109/TLT.2019.2913358

Ranalli, J. (2018). Automated written corrective feedback: how well can students make use of it? *Computer Assisted Language Learning*, *31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Ranalli, J., Feng, H.-H., & Chukharev-Hudilainen, E. (2018a). The affordances of process-tracing technologies for supporting L2 writing instruction. *Language Learning & Technology*, *23*(2), 1–11. https://doi.org/10125/44678

Ranalli, J., Feng, H.-H., & Chukharev-Hudilainen, E. (2018b). Exploring the potential of process-tracing technologies to support assessment for learning of L2 writing. *Assessing Writing*, *36*, 77–89. https://doi.org/10.1016/j.asw.2018.03.007

Rapp, C., & Kauf, P. (2018). Scaling academic writing instruction: Evaluation of a scaffolding tool (Thesis Writer). *International Journal of Artificial Intelligence in Education*, *28*(4), 590–615. https://doi.org/10.1007/s40593-017-0162-z

Révész, A., Kourtali, N.-E., & Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning*, *67*(1), 208–241. https://doi.org/10.1111/lang.12205

Rijlaarsdam, G., Janssen, T., Rietdijk, S., & Van Weijen, D. (2017). Reporting design principles for effective instruction of writing: Interventions as constructs. In R. Fidalgo, K. Harris, & M. Braaksma (Eds.), *Design principles for teaching effective writing* (Vol. 34, pp. 280–313). Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004270480_013

Rijlaarsdam, G., & Van den Bergh, H. (1996). The dynamics of composing—An agenda for research into an interactive compensatory model of writing: Many questions, some answers. In C. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 107–125). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Roeser, J., Torrance, M., & Baguley, T. (2019). Advance planning in written and spoken sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(11), 1983–2009. https://doi.org/10.1037/xlm0000685

Roeser, J., Torrance, M., De Maeyer, S., Leijten, M., & Van Waes, L. (2020). Analysing interkey intervals: Beyond means, medians and pause frequencies. To be presented at the *EARLI SIG online measurements*, Antwerp, Belgium.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/10.1002/widm.1075

Romero, C., & Ventura, S. (2019). Guest editorial: Special issue on early prediction and supporting of learning performance. *IEEE Transactions on Learning Technologies*, *12*(2), 145–147. https://doi.org/10.1109/TLT.2019.2908106

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, *34*, 39–59. https://doi.org/10.1016/j.compcom.2014.09.002

Roscoe, R. D., Jacovina, M. E., Allen, L. K., Johnson, A. C., & McNamara, D. S. (2016). Toward revision-sensitive feedback in automated writing evaluation. In *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 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. https://doi.org/10.1016/S0364-0213(82)80004-9

Salmeron-Majadas, S., Santos, O. C., & Boticario, J. G. (2014). An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context. *Procedia Computer Science*, *35*, 691–700. https://doi.org/10.1016/j.procs.2014.08.151

Santangelo, T., Harris, K., & Graham, S. (2016). Self-regulation and writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 174–193). New York, NY, US: Guilford Publications.

Scardamalia, M., & Bereiter, C. (1983). The development of evaluative, diagnostic and remedial capabilities in children's composing. In M. Martlew (Ed.), *The psychology of written language: Developmental and educational perspectives* (pp. 67–95). New York, NY, US: McMillan.

Schunk, D. H., & Swartz, C. W. (1993). Goals and progress feedback: Effects on self-efficacy and writing achievement. *Contemporary Educational Psychology*, *18*(3), 337–354. https://doi.org/10.1006/ceps.1993.1024

Sengupta, S. (2000). An investigation into the effects of revision strategy instruction on L2 secondary school learners. *System*, *28*(1), 97–113. https://doi.org/10.1016/S0346-251X(99)00063-9

Severinson-Eklundh, K., & Kollberg, P. (2001). Studying writers' revision patterns with S-notation analysis. In O. T. & L. C.M. (Eds.), *Contemporary tools and techniques for studying writing* (Vol. 10, pp. 89–104). Dordrecht, The Netherlands: Springer.

Shibani, A., Knight, S., & Shum, S. B. (2019). Contextualizable learning analytics design: A generic model and writing analytics evaluations. In *Proceedings of the Ninth International Conference on Learning Analytics & Knowledge* (pp. 210–219). New York, NY, US: Association for Computing Machinery. https://doi.org/10.1145/3303772.3303785

Sinharay, S., Zhang, M., & Deane, P. (2019). Prediction of essay scores from writing process and product features using data mining methods. *Applied Measurement in Education*, *32*(2), 116–137. https://doi.org/10.1080/08957347.2019.1577245

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, *57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Solé, I., Miras, M., Castells, N., Espino, S., & Minguela, M. (2013). Integrating information: an analysis of the processes involved and the products generated in a written synthesis task. *Written Communication*, *30*(1), 63–90. https://doi.org/10.1177/0741088312466532

Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College Composition and Communication*, *31*(4), 378–388. https://doi.org/10.2307/356588

Sonderlund, A., Hughes, E., & Smith, J. R. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, *50*(5), 2594–2618. https://doi.org/10.1111/bjet.12720

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, *12*(3), 175-200. https://doi.org/10.20982/tqmp.12.3.p175

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149–183. https://doi.org/10.1177/0741088316631527

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, *106*(2), 331–347. https://doi.org/10.1037/a0034752

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Stevenson, M., Schoonen, R., & de Glopper, K. (2006). Revising in two languages: A multidimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing*, *15*(3), 201–233. https://doi.org/10.1016/j.jslw.2006.06.002

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, *131*, 33–48. https://doi.org/10.1016/j.compedu.2018.12.005

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Tan, J. P.-L., Koh, E., Jonathan, C. R., & Yang, S. (2017). Learner dashboards a double-edged sword? Students' sense-making of a collaborative critical reading and learning analytics environment for fostering 21st century literacies. *Journal of Learning Analytics*, *4*(1), 117–140. https://doi.org/10.18608/jla.2017.41.7

Tappert, C. C., Villani, M., & Cha, S.-H. (2010). Keystroke biometric identification and authentication on long-text input. In L. Wang & X. Geng (Eds.), *Behavioral biometrics for human identification: Intelligent applications* (pp. 342–367). Hershey, PA, US: IGI Global. https://doi.org/10.4018/978-1-60566-725-6.ch016

Teasley, S. D. (2017). Student facing dashboards: One size fits all? *Technology, Knowledge and Learning*, *22*(3), 377–384. https://doi.org/10.1007/s10758-017-9314-3

Theelen, H., Willems, M., Van den Beemt, A., Conijn, R., & den Brok, P. (2019). Virtual internships in blended environments to prepare preservice teachers for the professional teaching context. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.12760

Thomas, R. C., Karahasanovic, A., & Kennedy, G. E. (2005). An investigation into keystroke latency metrics as an indicator of programming performance. In *Proceedings of the 7th Australasian Conference on Computing Education* (Vol. 42, pp. 127–134). Australian Computer Society, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1082440

Thorson, H. (2000). Using the computer to compare foreign and native language writing processes: A statistical and case study approach. *The Modern Language Journal*, *84*(2), 155–170. https://doi.org/10.1111/0026-7902.00059

Tillema, M., Van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2011). Relating self reports of writing behaviour and online task execution using a temporal model. *Metacognition and Learning*, *6*(3), 229–253. https://doi.org/10.1007/s11409-011-9072-x

Torrance, M., & Galbraith, D. (2006). The processing demands of writing. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 67–80). New York, NY, US: Guilford Publications.

Torrance, M., Thomas, G. V., & Robinson, E. J. (2000). Individual differences in undergraduate essay-writing strategies: A longitudinal study. *Higher Education*, *39*(2), 181–200. https://doi.org/10.1023/A:1003990432398

Vandermeulen, N., Leijten, M., & Van Waes, L. (2020). Reporting writing process feedback in the classroom: Using keystroke logging data to reflect on writing processes. *Journal of Writing Research*, *12*(1), 109-140. https://doi.org/10.17239/jowr-2020.12.01.05

Van Horenbeeck, E., Pauwaert, T., Van Waes, L., & Leijten, M. (2015). *A generic XML-structure for logging human computer interaction.* Retrieved from http://www.inputlog.net/wp-content/uploads/Generic_XML_structure_version-1_3.pdf (white paper)

Van Lier, L. (2000). From input to affordance: Social-interactive learning from an ecological perspective. In J. Lantolf (Ed.), *Sociocultural theory and second language learning* (Vol. 78, pp. 245–259). Oxford, England: Oxford University Press.

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. (2019). A multilingual copy task: Measuring typing and motor skills in writing with Inputlog. *Journal of Open Research Software*, *7*(30), 1-8. https://doi.org/10.5334/jors.234

Van Waes, L., Leijten, M., Wengelin, A., & Lindgren, E. (2012). Logging tools to study digital writing processes. In V. Berninger (Ed.), *Past, present, and future contributions of cognitive writing research to cognitive psychology* (pp. 507–533). New York, NY, US: Psychology Press.

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. https://doi.org/10.1016/S0378-2166(02)00121-2

Van Waes, L., Van Weijen, D., & Leijten, M. (2014). Learning to write in an online writing center: The effect of learning styles on the writing process. *Computers & Education*, *73*, 60–71. https://doi.org/10.1016/j.compedu.2013.12.009

Vehtari, A., Gelman, A., & Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint*. Retrieved from https://arxiv.org/abs/1507.02646

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, *57*(10), 1500–1509. https://doi.org/10.1177/0002764213479363

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Assche, F., Parra, G., & Klerkx, J. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing*, *18*(6), 1499–1514. https://doi.org/10.1007/s00779-013-0751-2

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wallot, S., & Grabowski, J. (2013). Typewriting dynamics: What distinguishes simple from complex writing tasks? *Ecological Psychology*, *25*(3), 267–280. https://doi.org/10.1080/10407413.2013.810512

Wang, Y.-J., Shang, H.-F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*, *26*(3), 234–257. https://doi.org/10.1080/09588221.2012.655300

Wengelin, Å. (2006). Examining pauses in writing: Theory, methods and empirical data. In K. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing: Methods and applications* (Vol. 18, pp. 107–130). Oxford, UK: Elsevier.

Wengelin, Å. (2007). The word-level focus in text production by adults with reading and writing difficulties. In M. Torrance, L. Van Waes, & D. Galbraith (Eds.), *Writing and cognition: Research and applications* (Vol. 20, pp. 67–82). Leiden, The Netherlands: Brill. https://doi.org/10.1163/9781849508223_006

Williamson, M. M., & Pence, P. (1989). Word processing and student writers. In B. Britton & S. Glynn (Eds.), *Computer writing environments: Theory, research, and design* (pp. 93–127). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Wilson, J. (2017). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing*, *30*(4), 691–718. https://doi.org/10.1007/s11145-016-9695-z

Wise, A. F., & Jung, Y. (2019). Teaching with analytics: Towards a situated model of instructional decision-making. *Journal of Learning Analytics*, *6*(2), 53–69. https://doi.org/10.18608/jla.2019.62.4

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. https://doi.org/10.1145/1188816.1188819

Wolsey, T. D., Lapp, D., & Fisher, D. (2012). Students' and teachers' perceptions: An inquiry into academic writing. *Journal of Adolescent & Adult Literacy*, *55*(8), 714–724. https://doi.org/10.1002/JAAL.00086

Woolner, P., Hall, E., Wall, K., & Dennison, D. (2007). Getting together to improve the school environment: User consultation, participatory design and student voice. *Improving Schools*, *10*(3), 233–248. https://doi.org/10.1177/1365480207077846

Woong Yun, G., & Park, S.-Y. (2011). Selective posting: Willingness to post a message online. *Journal of Computer-Mediated Communication*, *16*(2), 201–227. https://doi.org/10.1111/j.1083-6101.2010.01533.x

Xu, C. (2018). Understanding online revisions in L2 writing: A computer keystroke-log perspective. *System*, *78*, 104–114. https://doi.org/10.1016/j.system.2018.08.007

Xu, C., & Ding, Y. (2014). An exploratory study of pauses in computer-assisted EFL writing. *Language Learning & Technology*, *18*(3), 80–96. Retrieved from https://eric.ed.gov/?id=EJ1046527

Xue, H., & Hwa, R. (2014). Improved correction detection in revised esl sentences. In *Proceedings of the 52nd annual meeting of the association for computational linguistics: Short papers* (Vol. 2, pp. 599–604). https://doi.org/10.3115/v1/P14-2098

Yim, S., Wang, D., Olson, J., Vu, V., & Warschauer, M. (2017). Synchronous writing in the classroom: Undergraduates' collaborative practices and their impact on text quality, quantity, and style. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW'17)* (pp. 468–479). https://doi.org/10.1145/2998181.2998356

Zhang, F., Hwa, R., Litman, D., & Hashemi, H. B. (2016). Argrewrite: A web-based revision assistant for argumentative writings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 37–41). https://doi.org/10.18653/v1/N16-3008

Zhang, F., & Litman, D. (2015). Annotation and classification of argumentative writing revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 133–143). https://doi.org/10.3115/v1/W15-0616

Zhang, M., Hao, J., Li, C., & Deane, P. (2016). Classification of writing patterns using keystroke logs. In L. A. Van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society, Beijing, 2015* (pp. 299–314). Springer International Publishing. https://doi.org/10.1007/978-3-319-38759-8_23

Zhang, M., Zhu, M., Deane, P., & Guo, H. (2019). Identifying and comparing writing process patterns using keystroke logs. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology* (pp. 367–381). Springer International Publishing. https://doi.org/10.1007/978-3-030-01310-3_32

Zhu, M., Zhang, M., & Deane, P. (2019). Analysis of keystroke sequences in writing logs. *ETS Research Report Series*(1), 1–16. https://doi.org/10.1002/ets2.12247

Appendix A
Rubric academic summary

The grading rubric for the academic summary task as described in Chapter 4 is shown in Table A.1, with scores for all the five criteria. Points were subtracted for task compliance:

- No in-text citation: 0.5
- Between 200 words and 300 words: 0.5
- Less than 100 or more than 300 words: 1.0

This resulted in: Final grade = (total score rubrics / 2) - task compliance.

Table A.1: Grading rubric academic summary task

	1	2	3	4
Main idea	Main idea is not present.	Main idea is not clearly stated.	Main idea is mostly clear; or main idea is clear, but not within the first two sentences.	Main idea is clear and within the first two sentences.
Structure & organization	The paragraph is poorly structured and hard to follow.	The paragraph lacks some structure, and could be more cohesive and/or consistent.	The paragraph is well-structured, but could be more cohesive and/or consistent.	The paragraph is well-structured, cohesive, and consistent.
Content	More than one piece of critical supporting evidence is missing; or instances of incorrect material.	Critical supporting evidence is missing; or many instances of irrelevant material.	Most supporting evidence is included; some irrelevant material.	All supporting evidence is included; no incorrect or irrelevant material.
Language & paraphrasing	Author plagiarizes.	Author uses quotes or sentences too similar to the text.	Author uses own words, but limited variety in words and sentence structures.	Author uses own words, wide variety in words and sentence structures.
Grammar & mechanics	>5 spelling and/or grammar mistakes.	3–4 spelling and/or grammar mistakes.	1–2 spelling and/or grammar mistakes.	No spelling and/or grammar mistakes.

Appendix B
Annotation guidelines

This appendix provides the annotation guidelines for annotating revisions in the writing process, following the revision tagset as described in Chapter 5.

Introduction

During writing, a large variety of revisions are made. By distinguishing these different types of revision during the writing process, we could gain an earlier and more precise identification of writing abilities and difficulties. This could be used to design personalized and real-time feedback or instruction targeting these specific writing difficulties.

With this annotation scheme, we aim to categorize different properties of revision, which enables us to distinguish between different types of revision. The annotation scheme below lists all properties which cannot be (easily) automatically extracted from writing time course data. Therefore, these annotation guidelines were created to guide the manual annotation of these properties.

Data prerequisites

To be able to start annotating revisions during the writing process, enriched writing time course data is needed. From the raw time course data (also known as keystroke log data), the following features need to be extracted:

1. Revision events: Chunks of keystrokes which are ended by a revision or by text production at another location in the writing product. For each revision event we need: the revision number (unique identifier; RID), the removed characters, and the typed characters.
2. Replay of the writing process (if possible, with eye tracking data).

Start annotation with CyWrite

1. Open the blank annotation spreadsheet.
2. Save a copy of the annotation spreadsheet with the name: [annotator name]_[task], e.g., john_pilot.xlsx.
3. Type your name in cell Z3.
4. Pick the file you would like to annotate, using the filter function in cell Y1.
5. Change the font of columns B, C, E into a serif typeface you feel comfortable with.
6. Open the corresponding file in CyWrite by clicking on the URL.
7. Begin filling out the annotation spreadsheet according to the rules outlined below.
8. Once finished, pick a new file to annotate and repeat steps 3–7.

Revision properties for manual annotation

The revision properties listed below only consist of the properties which cannot be automatically annotated. This list thus cannot be considered as a complete list of properties of revisions. Since not every chunk would be a revision, we first need to annotate whether the chunk is a revision and where the revision ended. Most annotations are binary, where 0 = **'no'** and 1 = **'yes'**.

General

1. Revision [Y/N]. The chunk is a revision, unless it is just fluent text production at the leading edge (cursor location). Values: 0,1. For an example see Table B.1. Thereafter, every chunk that is coded as a revision needs to be annotated with each of the following properties.

Table B.1: Annotation example for revision [Y/N]

RID	Removed characters	Typed characters	Revision	Revision end
1		to	0	
2	to	To hear myself	1	To/ hear myself

2. Revision end. All characters typed up to where the revision ended and text production started. If the writer only revised a character within a word, the revision ends at the end of that word. If the writer revised a word/phrase/sentence to follow a new train of thought,

Table B.2: Annotation example for revision end

RID	Removed characters	Typed characters	Revision	Revision end
1	I wriet	I write every day	1	I write/ every day
2	every day	novels and poems/.	1	novels and poems!./

be lenient: put the revision end as far as possible. If there is already a slash in revision end (because the writer typed a slash), please change the typed slash into an exclamation mark (!). Values: put a slash where the revision ended (see Table B.2).

C. Orientation

The aspects of the text the revision is oriented towards. If it is unclear which orientation applies, pick all possible orientations. If something could be a semantic change, but you are not sure, for example, when only a few characters are replaced, you could pick both surface and semantic change. However, based on your ideas about the expertise/vocabulary of the writer, you might sometimes be more conservative. For example, changing 'interpersonally' into 'as interpersonally' would be a semantic change for a native writer, while it could be a grammar change for a second language writer who was struggling with the language earlier on in the text as well. Values: C.1. Surface; C.2. Deep/semantic.

C.1. Surface

Revisions that are conventional copy-editing operations (C.1.1–C.1.7) or revisions that paraphrases a concept in the text, but do not alter it (C.1.8). Within this subcategory pick typography first, then grammar, followed by spelling. So, do not select spelling when it was only a typography revision. If the correct subcategory is unclear, pick all possible orientations. Values: C.1.1. Typography (t); C.1.2. Capitalization (C; automatically filled-in); C.1.3. Punctuation (p; automatically filled-in); C.1.4. Spelling (s); C.1.5. Grammar (g); C.1.6. Cosmetics/presentation (c); C.1.7. No change (n; automatically filled-in); C.1.8. Wording/phrasing (w). For examples, see Table B.3.

C.1.1. Typography (t). A revision as a result of a typographic error, a slip of the finger, or unintended keystrokes. These are keystrokes that are often quickly revised, often happen in adjacent keys on the keyboard, or are transpositions of two adjacent characters.

Table B.3: Annotation example for surface type

RID	Removed chars.	Typed chars.	Revision	Surface change	Surface type
1	barin	brain	1	1	t
2		lisning music	0		
3		t [listning]	1	1	s
4	t [lisning]		1	1	s
5		is river	0		
6		a [is a river]	1	1	g
7	, howev	. However,	1	1	g
8	has the ability to	has the power to	1	1	w

C.1.4. Spelling (s). A revision aimed at revising the spelling of the target, for example, if a space is added/removed to concatenate or separate two words. A revision is not aimed at spelling, when, based on your ideas about the expertise/vocabulary of the writer, you would expect the writer to know how to spell this word, for example, because the writer has written it correctly before or because the writer has correctly spelled words with similar spelling rules.

C.1.5. Grammar (g). A revision aimed at revising the grammar of the target or punctuation outside a word (which is not yet coded as punctuation), for example, if a writer is debating between ending the sentence or not. Examples:

C.1.6. Cosmetics/presentation (c). A revision aimed at changing the visual presentation of the text, but which are not aimed at structuring the text. For example, changes in font size, italics.

C.1.8. Wording/phrasing (w). A revision made to paraphrase a concept, but not to alter the concept.

C.2. Semantic (deep)

Revisions that change the semantics or meaning of the text.

C.2.1. Microstructure changes. Revisions that change the semantics of the text but do not affect the summary of the text (C2.1). Select the type of semantic revision from the drop-down menu (you can type the first letter(s) and hit enter).

- **C.2.1.1. Supporting info**. A revision aimed at changing information, examples, or explanations.

- **C.2.1.2. Emphasis**. A revision aimed at changing the emphasis on a line of reasoning or findings.
- **C.2.1.3. Understate**. A revision aimed at changing the understatement on a line of reasoning or findings.
- **C.2.1.4. Coherence**. A revision aimed at changing the way the text is tied together by linguistic devices, such as additionally, however, etc.
- **C.2.1.5. Cohesiveness**. A revision aimed at changing the organization of the text or the flow of the sentences or paragraphs. For example, by starting reordering paragraphs or creating section headers.

C.2.2. Macrostructure changes Revisions that change the semantics or meaning of the text and affect the summary of the text. Select the type of semantic revision from the drop-down menu (you can type the first letter(s) and hit enter).

- **C.2.2.1 Overall aim**. A major revision that alters the overall aim of the text.
- **C.2.2.2 Subtopic**. A major revision that alters a subtopic of the text.

D. Evaluation

The evaluation of a revision, only necessary for typography, spelling, and grammar revisions (C.1.1, C.1.4, C.1.5). In all other cases, it is auto-completed with *NA*. Also type *NA* if the revision includes punctuation outside words, and it could be both correct as well as incorrect or it is impossible to evaluate (e.g., shortening a sentence by changing a comma into a period). Values: 0,1, *NA* (see Table B.4).

D.1. Correct start. Revision occurred within a linguistic domain with correct spelling and grammar (for the targeted part of the word/sentence).

D.2. Correct finish. Revision resulted in the correct spelling/grammar for the targeted part of the word/sentence (up to where the revision ended, see General, 2).

Table B.4: Annotation example for evaluation

RID	Removed chars.	Typed chars.	Revision	Correct start	Correct finish
1	sien	scientifiz	1	0	1
2	iz	ic	1	0	1
3	ic	ik	1	1	0

F. Linguistic domain

Domain in which the revision is made; pick the smallest possible domain (smallest number). This is the domain which is affected by the revision. Select the domain from the drop-down menu (you can type the first letter(s) and hit enter).

- **F.1.1. Subword**. Revision affects a single (sequence of) characters but not a complete word. This also includes spaces and punctuation.
- **F.1.2. Word**. Revision affects one or more words and 0 or more non-alphanumeric characters.
- **F.1.3. Phrase**. Revision affects one or more phrases (a group of words belonging to each other or having the same grammatical function inside the sentence)
- **F.1.4. Clause**. Revision affects one or more clauses (a sentence part that can express a complete composition).
- **F.1.5. Sentence**. Revision affects one or more sentences.
- **F.1.6. Paragraph**. Revision affects one or more paragraphs.

G. Spatial location

The position in the text where the revision started. Examples for all features related to spatial location described below are shown in Table B.5.

G.1. Word finished. The word was finished when the writer started deleting characters, or the insertion happened outside word boundaries. A word is considered finished if in the context of the previous words, no more characters were expected for the current word. Thus, either the last character of the word was typed, or a character was typed that was a likely typo for the last character of the word. The word (1) does not have to be spelled correctly (e.g., in 'he aets' the word is finished, and the expected word is 'eats'), and (2) a correctly spelled is not always a finished word (e.g., in 'he say' the word is not yet finished, you would expect 'says'). If, from the context, it is unclear whether the word is finished, type *NA*. Values: 0,1, *NA*.

G.2. Intended word. If the word was not finished (D.1 = 0), specify what you think the intended word could have been, based on the previous words, and with the correct spelling and grammar. If the intended word cannot be guessed, type *NA*. Values: intended word, *NA*.

G.3.1. Word initial. Characters are removed up to the start of a word, or characters are inserted from the start of a word. Values: 0,1.

G.3.2. Clause initial. Characters are removed up to start of the clause or inserted from start of the clause. Also: If characters are removed up to halfway through the first word of a clause, and one or a few characters are reused to type a new clause, it is still a clause initial. Punctuation or conjunction (e.g., that, who, but) at clause boundaries are included as clause initial. For example, in: "I like walking, but I prefer biking", revisions would be considered clause initial if they are revised (1) up to 'but', or (2) up to and including 'but'. Values: 0,1.

G.3.3. Sentence initial. Characters are removed up to the start of the clause or inserted from the start of the sentence. Also: If characters are removed up to halfway the first word of a sentence, and one or a few characters are reused to type a new sentence, it is still a clause initial. Punctuation at sentence boundaries are included as sentence initial. Values: 0,1.

Table B.5: Annotation example for spatial location

RID	Removed	Typed	Rev.	Word finished	Intended word	Word initial	Clause initial	Sentence initial
1	poeple	people	1	1	0	1	0	0
2	Poe	People	1	0	people	1	1	1
3		What	0					
4	at	en [*When*] I eat.	1	1	0	0	1	1
5	.	, I eat fish	1	1	0	0	1	0

J. Sequencing

Whether the revision was caused by or related to the previous revision. Only consider the closest previous revision (not earlier revisions). Examples for all features related to sequencing described below are shown in Table B.6.

J.1. Overrides previous. A revision that overrides the previous revision; it is repetitive at cursor location (repetition of changing the same linguistic domain over and over again); or a revision of a previous revision that had a similar target (RID 3, Table B.6). Values: 0,1.

J.2. Continues previous. A revision continues on a previous revision (not the previous text production); the revision is caused by a previous revision. This can happen when:

- There are multiple typographic revisions within the same word but at a different target in the word (otherwise it would be overrides; RID 6, Table B.6);
- A period is replaced by a comma, so the sentence-initial capital letter needs to become lowercase (RID 9, Table B.6);
- A phrase is inserted but the insertion is interrupted by a typo (RID 12, Table B.6). You would have put the revision end further in the text.

Values: 0,1.

Table B.6: Annotation example for sequencing

RID	Removed chars.	Typed characters	Revision	Overrides	Continues
1		sciense	0		
2	se	e	1	0	0
3	e	ce.	1	1	0
4		Sz	0		
5	z	cienze	1	0	0
6	ze	ce.	1	0	1
7		I am	0		
8	I	We	1	0	0
9	am	are	1	0	1
10		According to research	0		
11	research	some resae	1	0	0
12	ae	earch studies	1	0	1

Appendix C
Revision visualizations step 1

The following twenty visualizations were presented for the round table sessions with the writing researchers, as first step into the design of the revision dashboard (see Chapter 7).

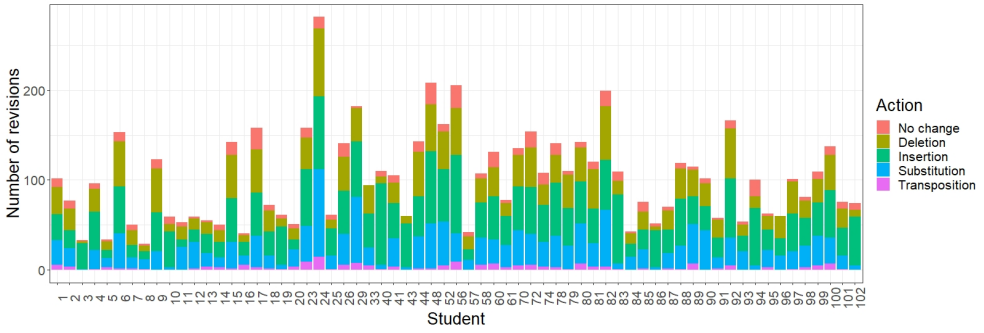

Figure C.1: Frequencies of revision actions.

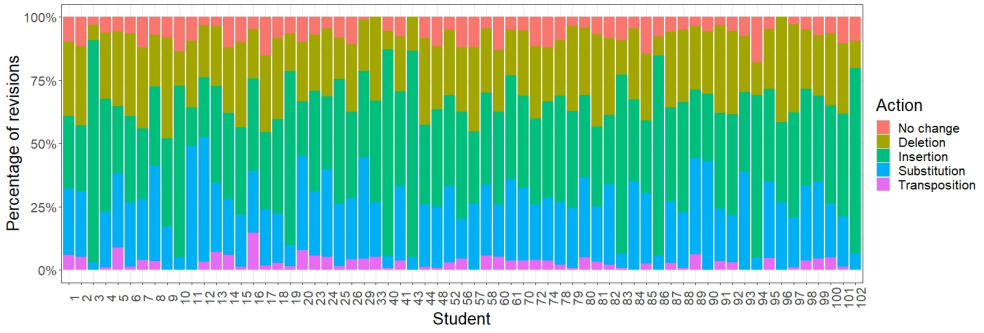

Figure C.2: Percentages of revision actions.

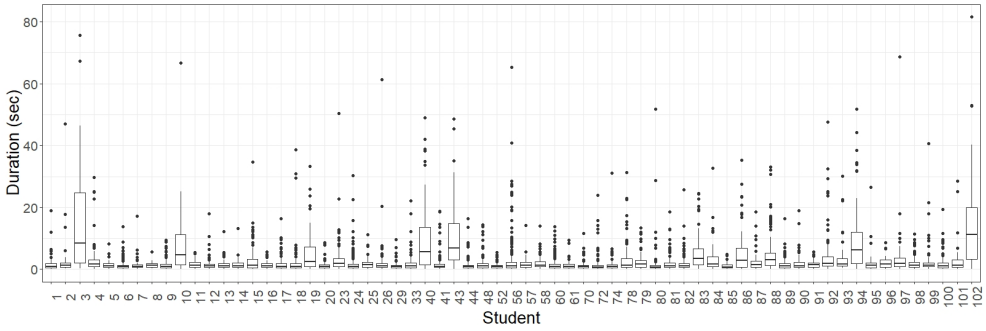

Figure C.3: Boxplot of revision durations.

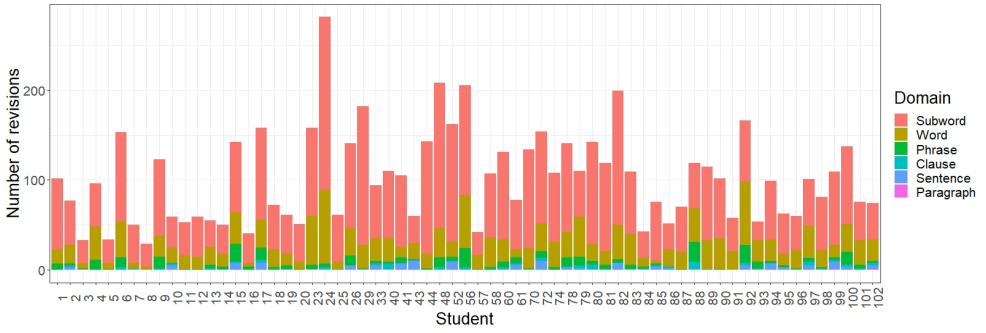

Figure C.4: Frequencies of revision domains.

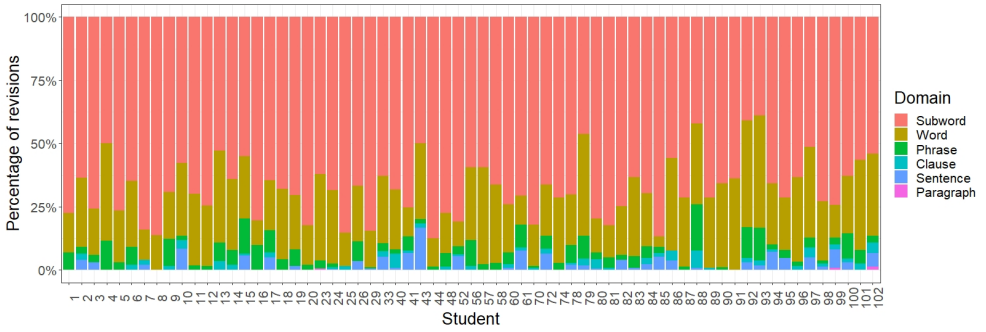

Figure C.5: Percentages of revision domains.

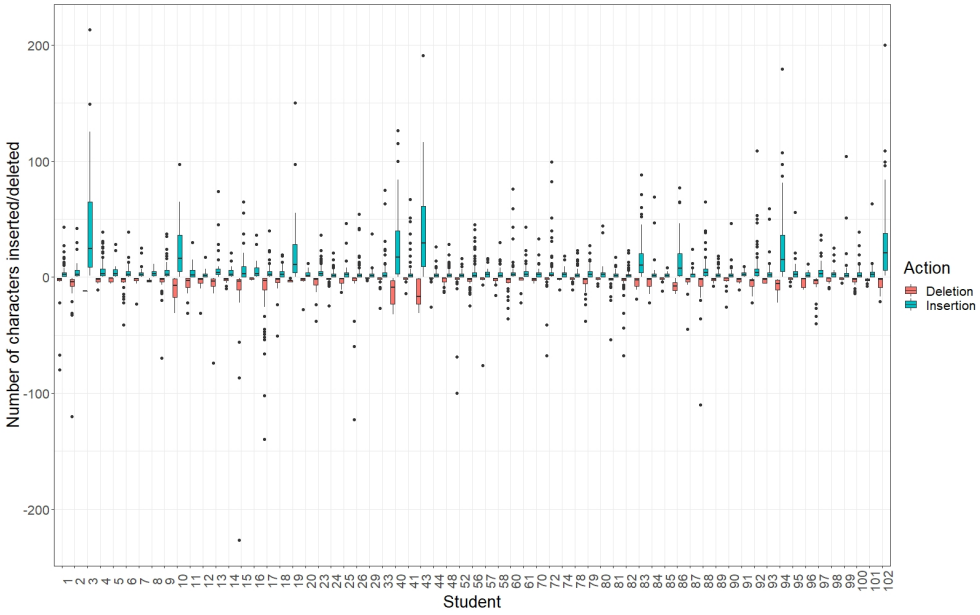

Figure C.6: Boxplot of revision domains.

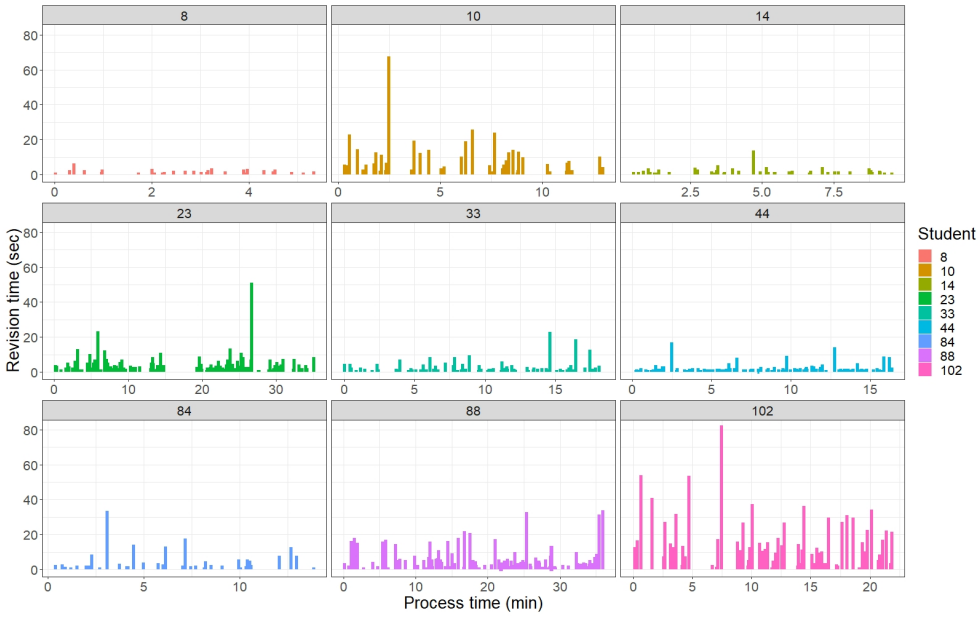

Figure C.7: Revision duration per temporal location.

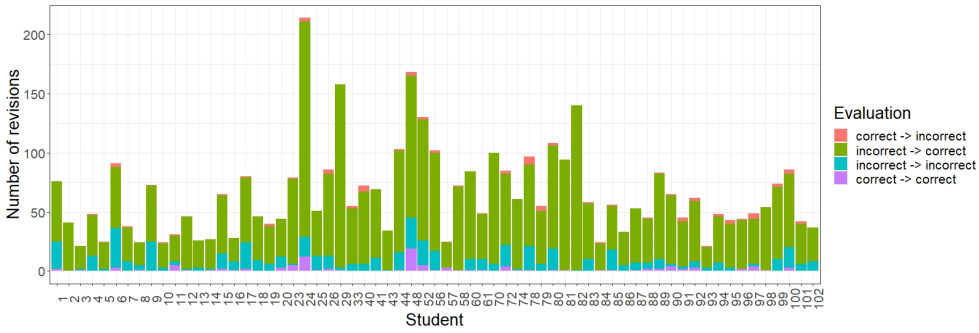

Figure C.8: Frequencies of revision evaluations.

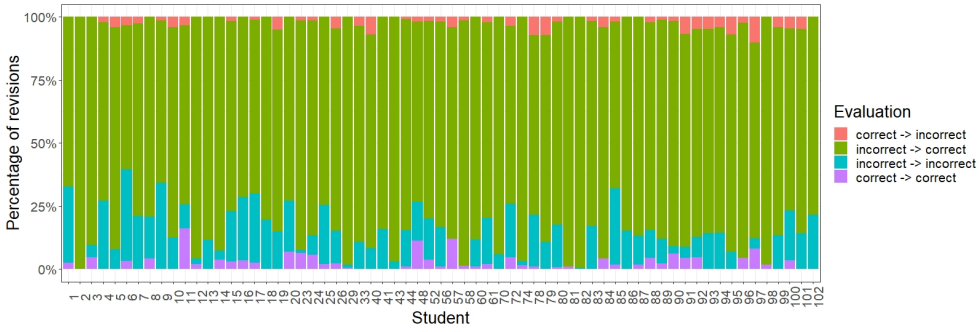

Figure C.9: Percentages of revision evaluations.

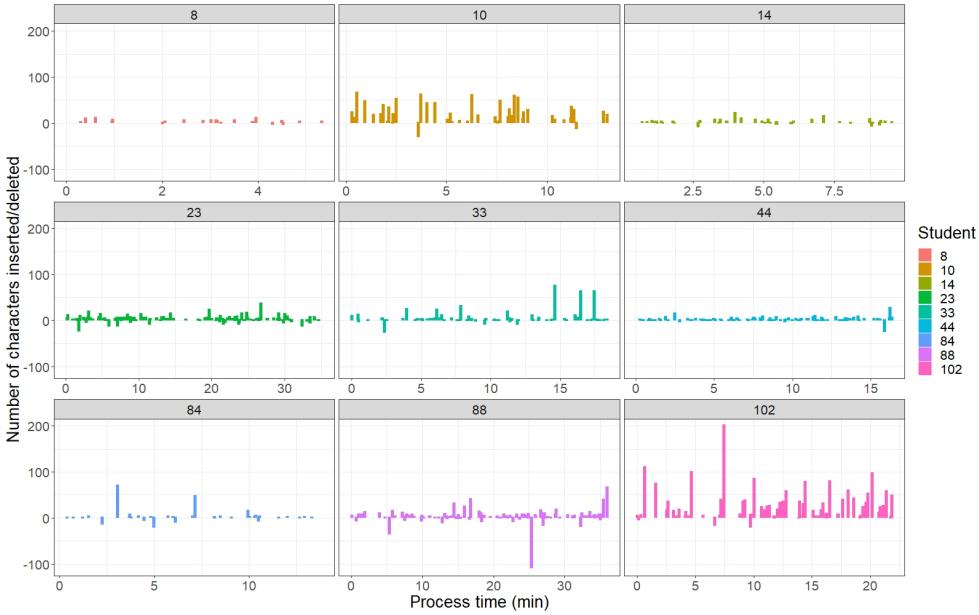

Figure C.10: Revision domain per temporal location.

Figure C.11: Revision duration per temporal location and orientation.

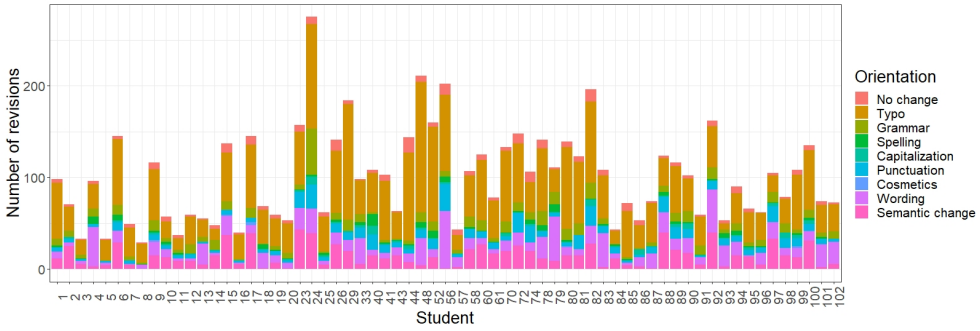

Figure C.12: Frequencies of revision orientations.

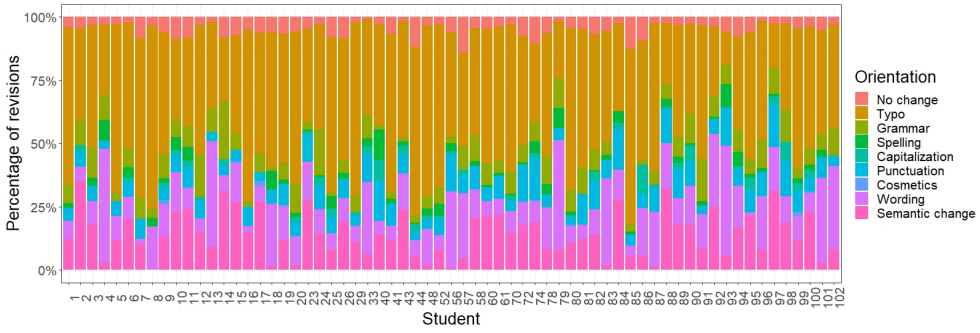

Figure C.13: Percentages of revision orientations.

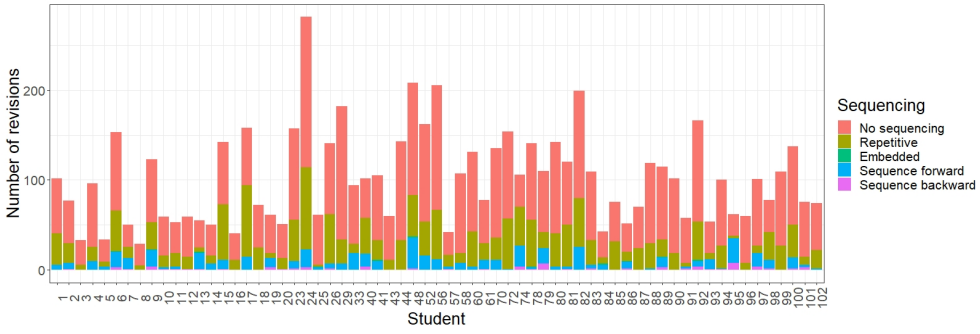

Figure C.14: Frequencies of revision sequencing.

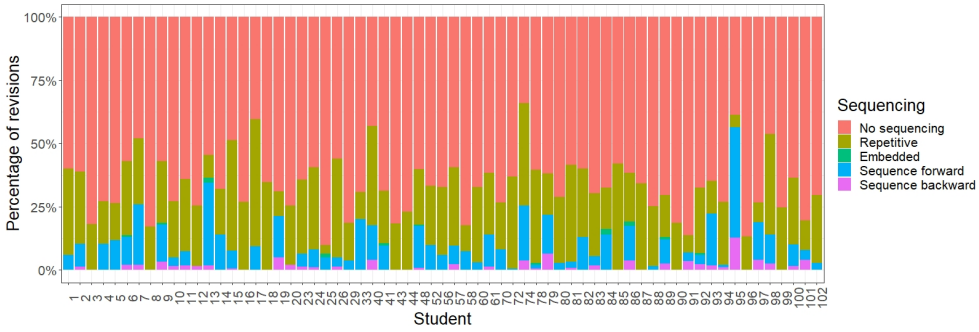

Figure C.15: Percentages of revision sequencing.

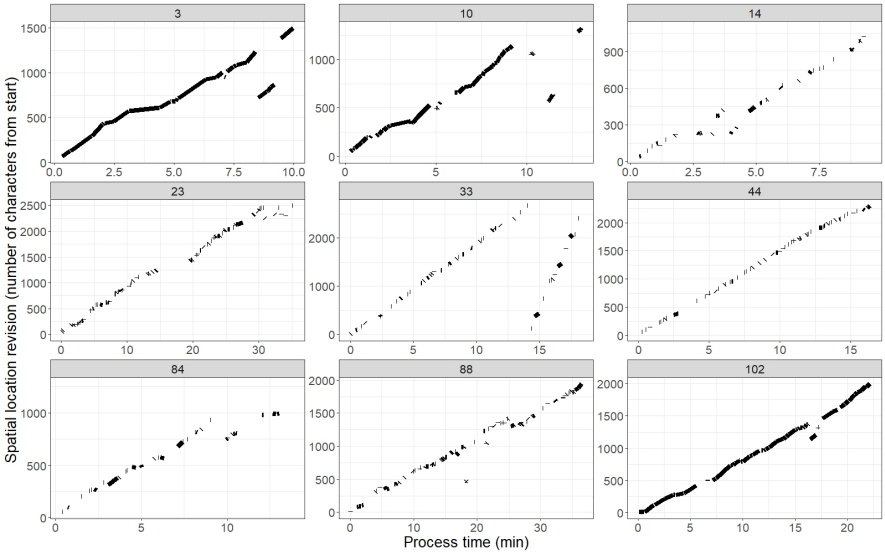

Figure C.16: Revision spatial location per temporal location.

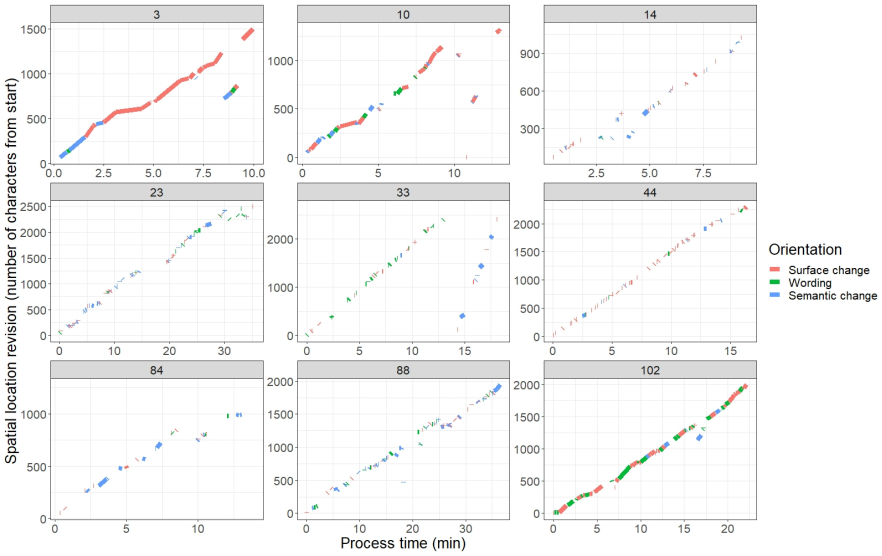

Figure C.17: Revision spatial location per temporal location and orientation.

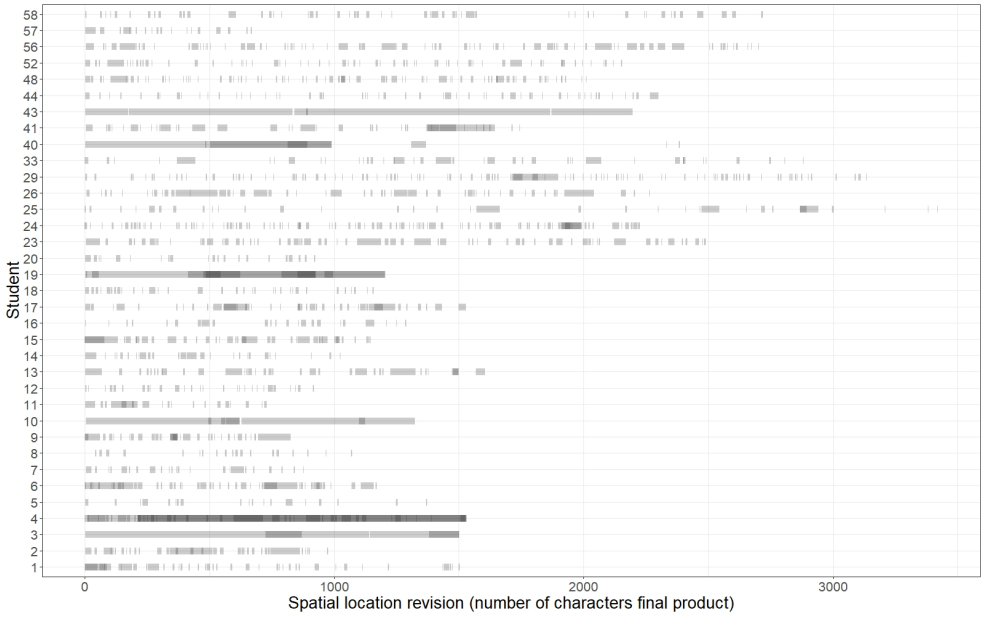

Figure C.18: Revision spatial location and sequencing.

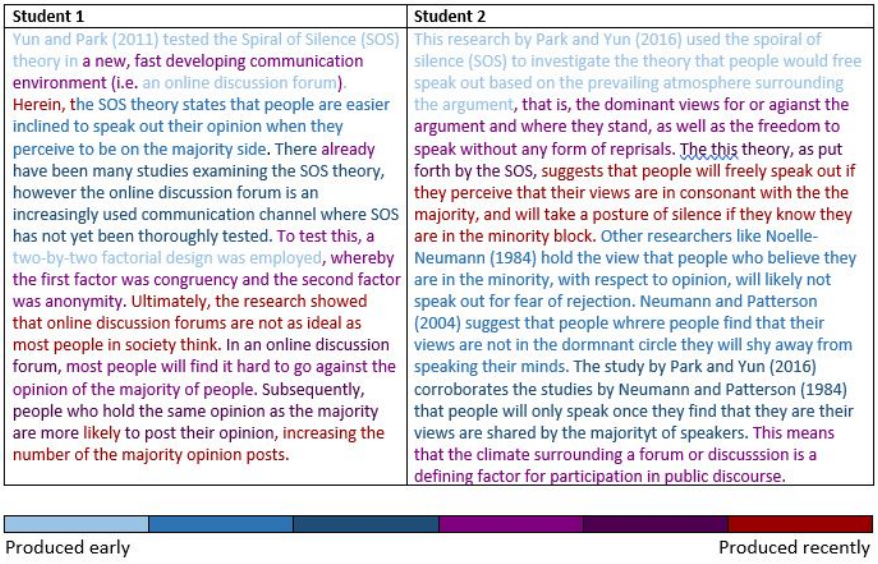

Figure C.19: Revision temporal location and spatial location in writing product.

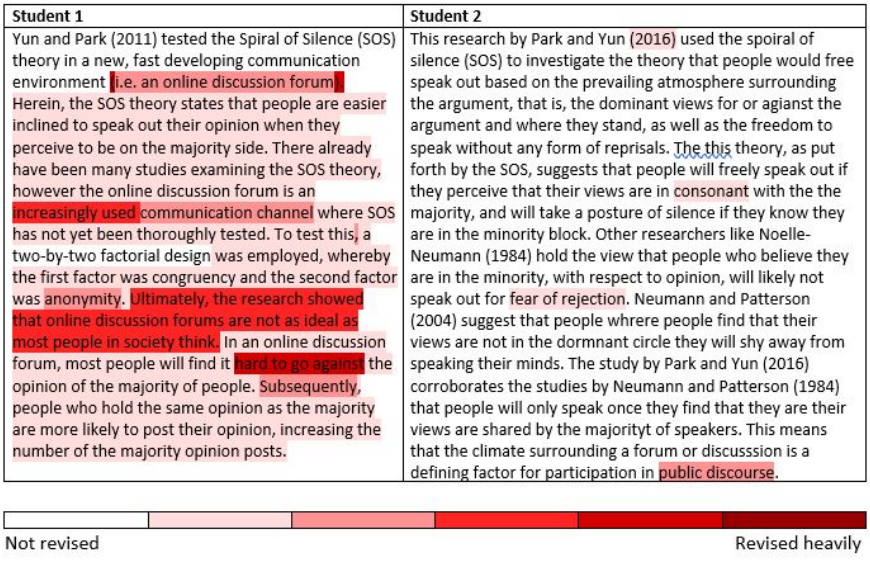

Figure C.20: Revision spatial location and sequencing in writing product.

Appendix D
Revision visualizations step 2

The following nine visualizations were presented for the focus session with the writing teachers, as second step into the design of the revision dashboard (see Chapter 7).

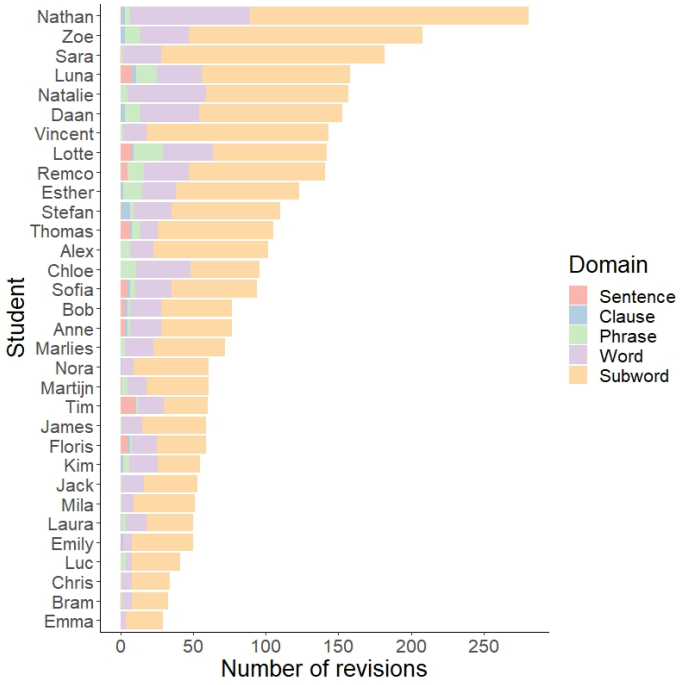

Figure D.1: Size of the revisions: domain that was affected.

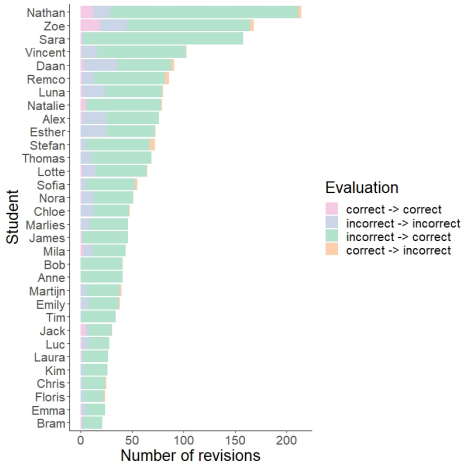

Figure D.2: Evaluation of the revisions.

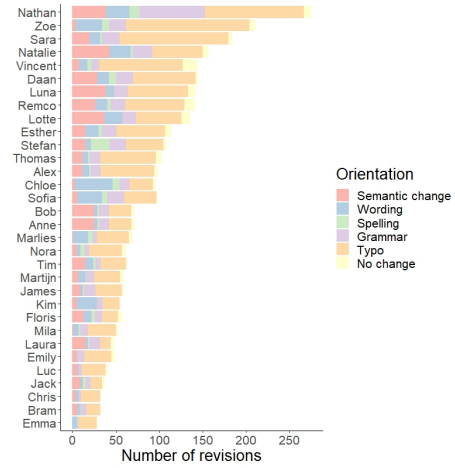

Figure D.3: Orientation of the revisions.

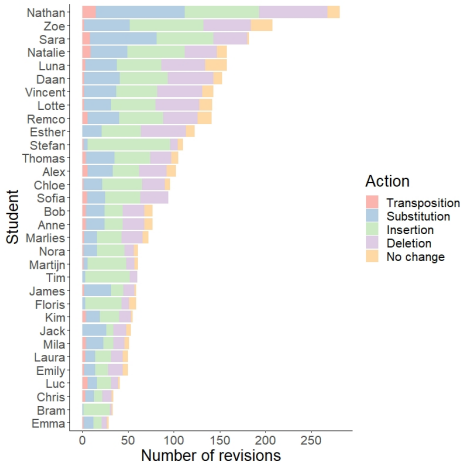

Figure D.4: Action of the revisions: operation made.

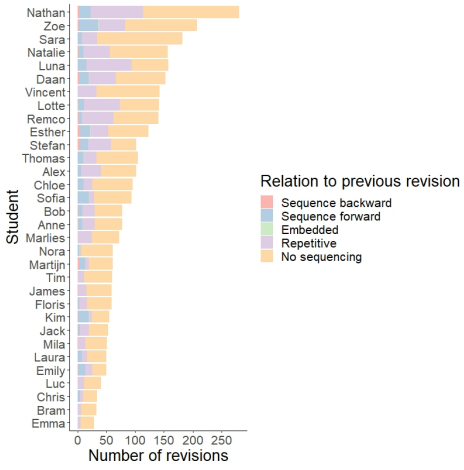

Figure D.5: Sequencing: relationship between subsequent revisions.

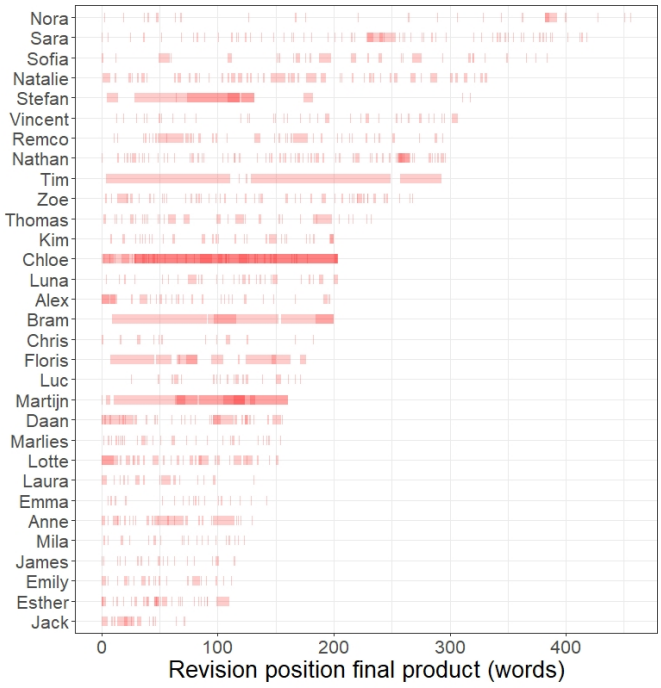

Figure D.6: Density and spatial location of the revisions.

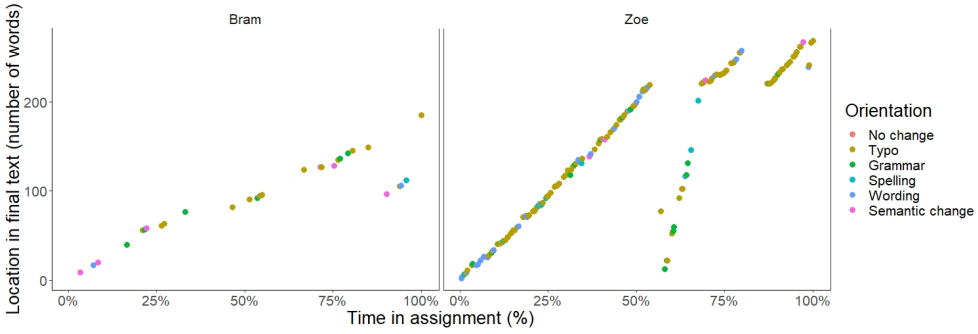

Figure D.7: Spatial location, temporal location, and orientation of the revisions.

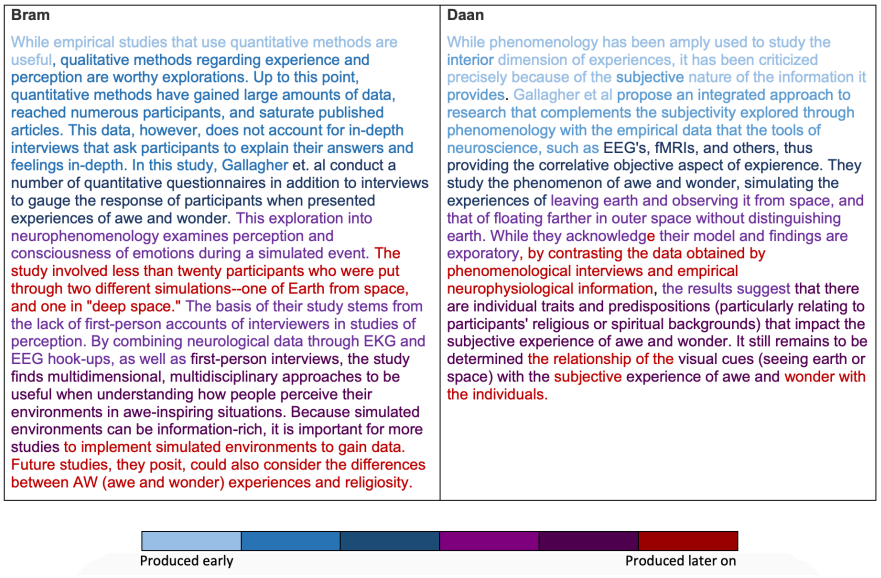

Figure D.8: Temporal location of the revisions in the final text.

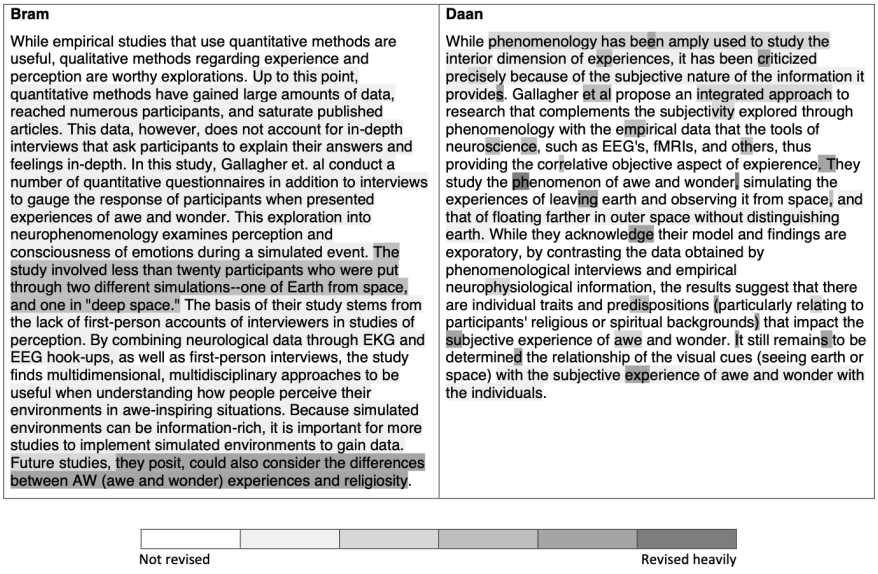

Figure D.9: Spatial location of the revisions in final text.

Summary

Writing plays an important role in higher education. However, several studies showed that students have difficulties with creating academic texts and teachers often complain about students' poor writing skills. Insight into students' cognitive and behavioral actions involved in writing; their writing processes, allows for a better understanding of when, where, and why students struggle. This knowledge could in turn be used for feedback and instruction to improve students' writing. For example, insight into the writing process could enhance students' awareness of their writing progress, and thereby improve effective development of task strategies as well as students' ability to self-regulate their writing.

Unfortunately, it is often difficult or even impossible for teachers to gain access to students' writing processes, especially in large classrooms or online settings. One unobtrusive and scalable solution to capture students' writing processes is the use of keystroke logging. With keystroke logging, every key pressed on a keyboard during writing is recorded, resulting in a detailed and timed overview of each key typed by a student. Literature has shown that these keystroke data can provide some insight into students' writing processes.

However, there is a large gap between these fine-grained keystrokes and the higher-level writing processes. In addition, it is still largely unknown how these detailed keystroke data can be used to provide teachers with meaningful insight into the writing process. Therefore, the current dissertation aims to identify how keystroke logging can be used to gain meaningful insight into students' writing processes. This is done in four steps: (1) identifying stakeholders' needs, (2) determining capabilities of keystroke analysis, (3) gaining insights from keystroke data, and (4) operationalizing insights from keystroke data. These four steps are divided over six studies detailed in six chapters.

Identifying stakeholders' needs

In **Chapter 2** we identified the stakeholders' needs. Although keystroke logging can provide a multitude of metrics about students' writing, little is known about the critical indicators for providing automated feedback. Therefore, this chapter describes a participa-

tory approach, to identify the indicators of students' writing processes that are meaningful for educational stakeholders in order to provide instruction and/or feedback on students' writing processes. In total, five participatory sessions were held with five distinct groups of stakeholders: bachelor and postgraduate students, teachers, writing specialists, and professional development staff. The results showed that all stakeholders desired indicators on all the main processes in writing: planning, translating, reviewing, and monitoring processes. Indicators included, for example, information on students' planning strategies, how students use evidence in their writing, the depth of revisions, and students' understanding of the task. It was considered important to extract both behavioral indicators (e.g., speed of writing), as well as higher-level cognitive indicators (e.g., critical thinking) and behavioral indicators in relation to time (e.g., spread of revisions). These findings prioritize which indicators need to be extracted. In addition, it was shown that the levels at which these indicators were discussed as well as the terminology used, differed between the groups of stakeholders. This further highlights the need for human-centered, participatory approaches to design and develop writing analytics tools.

Determining capabilities of keystroke analysis

After identifying the stakeholders' needs, we need to determine what is technically feasible, given the keystroke data available. To better understand how keystroke data map to higher-level cognitive writing processes, in **Chapter 3** we investigated the sensitivity of frequently used keystroke features across tasks with different cognitive demands. Two keystroke datasets were analyzed: one consisting of a copy task and an email writing task, and one with a larger difference in cognitive demand: a copy task and an academic summary task. The differences across tasks were modeled using Bayesian linear mixed effects models. Posterior distributions were used to compare the strength and direction of the task effects across features and datasets. The results showed that the mean of all interkeystroke intervals were found to be stable across tasks. Features related to the time between words and (sub)sentences only differed between the copy and the academic task. Lastly, keystroke features related to the number of words, revisions, and total time, differed across tasks in both datasets. To conclude, our results indicate that the latter features are related to cognitive load or task complexity, and hence might be used to gain insight into students' writing processes. In addition, the findings show that keystroke features are sensitive to small differences in the writing tasks at hand.

For providing teachers with (automated) insights in students' writing processes, a first step is to identify students at risk during writing. Therefore, in **Chapter 4** we determine the relation between the keystroke data and writing quality. For timely feedback, we would like to be able to identify students at risk as soon as possible. Although literature has shown that keystroke data can be used to predict writing quality *after* a draft has been submitted, this chapter investigated if keystroke data can also be used to predict writing quality *during* the writing process. Keystroke data were analyzed from 126 English as a second language learners performing a timed, academic summarization task. Writing quality was measured by final grade. Based on previous literature, 54 keystroke features were extracted. Machine learning models (regression and classification) were used to predict final grade and classify students at risk at several points during the writing process. In contrast to previous work, none of the models were able to outperform the baseline in predicting final grade (regression) after the full writing process was finished. In addition, the models for predicting pass/fail (classification) only slightly outperformed the baseline. Lastly, the relation between the writing process features and writing quality changed over time during the writing process. This stresses the importance to analyze the keystroke features in relation to time. Moreover, this shows that the relation between keystroke data and writing quality is rather limited. Therefore, in the remainder of the dissertation the focus shifted from using keystroke data to model writing *processes* as opposed to writing *quality*.

Gaining insights

With the insights in the stakeholders' needs and the possibilities and limitations of the keystroke data, we then turn to modeling students' writing processes. Here, we scope the dissertation to specifically focus on revision processes only, as stakeholders indicated this is one of the most desired processes to gain insight into, that is rather directly observable in the keystroke data.

In **Chapter 5** we provided a methodological approach to studying revisions in depth. Given the importance of revision in writing, revision has been a main topic of interest in writing research. Several models of revision have been developed, and a variety of taxonomies have been used to measure revision in empirical studies. Current advances in data collection and analysis have made it possible to study revision in more detail. However, a specific approach of how to do this was lacking. Therefore, the chapter provides a comprehensive product-oriented and process-oriented tagset of revisions. The presented tagset

consists of ten properties of revisions: processing, trigger, orientation, evaluation, action, linguistic domain, spatial location, temporal location, duration, and sequencing. We described how the features related to these properties can be extracted manually or automatically, using keystroke logging, screen replays, and eye tracking. As a proof of concept, we showed that this tagset could be used to annotate revisions made by higher education students with various backgrounds in various academic tasks. This tagset forms the basis for studying and visualizing revisions in the following chapters.

Based on the tagset, in **Chapter 6**, we further explored one specific type of revision: the revision of typographic errors (slips of the finger). These revisions can have a large influence on the writing process, and hence also on the analysis of the writing process. On the one hand, these types of revisions are low-level, and hence less important, so it might be advisable to ignore them for certain research questions. On the other hand, it is important to identify these revisions, as they can (unwillingly) break the flow in writing. Therefore, this chapter describes a process-based model to automatically identify typographic errors and their revisions. First, typographic errors and their revisions are characterized based on temporal and bigram properties extracted from keystroke data. Thereafter, a process-based model is trained on typographic error revisions on a copy task dataset to automatically identify these revisions. Finally, this model is evaluated in a more natural setting: a regular (source-based) writing task. Results show that it is possible to identify typographic errors using keystroke data only, especially in a copy task. However, the results did not generalize well to the source-based writing task. Hence, for the modeling of typographic error revisions, we need to further improve the accuracy of the machine learning models.

Operationalizing insights

Lastly, with the insights obtained from modeling the keystroke data we return to the stakeholders. In **Chapter 7** we determined how these models can be presented and integrated into the learning design, to ultimately improve the learning and teaching of writing. Data about students' learning processes are often displayed in so-called learning dashboards. Yet, simply providing teachers with data on students' learning processes is not necessarily beneficial for improving learning and teaching; the data need to be *actionable*. Recently, human-centered learning analytics has been suggested as a solution to realize more effective and actionable dashboards. Accordingly, we used this human-centered approach to design an *interpretable* and *actionable* dashboard to provide insight into students' revision processes.

The design consisted of three iterative steps. First, visualizations on students' revision processes, created from keystroke data, were evaluated with writing researchers. Second, focus group session with teachers of academic writing were used to co-design a paper prototype of the dashboard using the updated visualizations. Finally, the paper prototype was transformed into a digital prototype and evaluated by other teachers in individual user tests combined with interviews. The results showed that this approach was useful for designing an interpretable dashboard. The teachers envisioned a wide variety of actions, ranging from individual coaching to classroom-wide instruction. Hence, the dashboard opens new perspective to bring writing analytics to the classroom, by inviting students to reflect on concrete representations of their writing processes.

Discussion and Conclusion

Based on the primary focus on keystroke logging and the variety of methods employed in each of the studies, several reflections on the use of keystroke logging educational practices as well as in writing analytics research can be drawn. For the use of keystroke logging in educational practice, an adaptive approach is necessary, as there is no one-size-fits-all approach for improving students' writing processes. In addition, an appropriate technical infrastructure needs to be provided, which ensures data privacy and transparency, alleviates the burden for teachers, and requires little data literacy. For the use of keystroke logging in writing research, several opportunities for future work are identified. First, the data-driven approaches (as used in this dissertation) can be used to for more scalable analyses (e.g., as compared to manual annotation), can be used to identify points of interest in the keystroke data, and shed light on constructs that are hard to theoretically define. Second, given how keystroke data change over time during the writing process, it is important to also consider the temporal aspects of keystroke data. Lastly, given the sensitive keystroke data, it is important to combine keystroke features and to apply specific types of machine learning algorithms or more advanced statistical methodologies. This will result in more robust findings that are easier to relate to underlying cognitive processes involved in writing.

To conclude, we showed that to gain meaningful insight into students' writing processes, it is important to consider the technical possibilities of the keystroke data as well as the stakeholders' needs. By using data-driven approaches, automatic and scalable insights can be gained into students' revision processes. By using a human-centered approach, these insights can be transformed into *meaningful* insights with educational applications.

Samenvatting (Dutch Summary)

Schrijven speelt een belangrijke rol in het hoger onderwijs. Uit verschillende onderzoeken is echter gebleken dat studenten vaak moeite hebben met het schrijven van academische teksten en docenten vaak klagen over de slechte schrijfvaardigheid van de studenten. Inzicht in de cognitieve processen en gedragingen van studenten tijdens het schrijven, hun schrijfprocessen, kan zorgen voor een beter begrip over waar, wanneer en waarom studenten moeite hebben met het schrijven. Deze kennis kan vervolgens worden gebruikt voor instructie en terugkoppeling om het schrijfproces te verbeteren. Inzicht in het schrijfproces zou bijvoorbeeld het bewustzijn van studenten over hun schrijfvoortgang kunnen vergroten en kan daardoor effectief de ontwikkeling van taakstrategieën en zelfregulatie verbeteren.

Helaas is het voor docenten vaak moeilijk en soms zelfs onmogelijk om toegang te krijgen tot het schrijfproces van studenten. Recent zijn echter observatiemethodes ontwikkeld die een schaalbare oplossing hiervoor bieden en bovendien het schrijfproces zelf niet verstoren. Het betreft meer bepaald toetsregistratieprogramma's. Bij het verzamelen van toetsaanslagen wordt de tijd van elke toets die tijdens het schrijven op een toetsenbord wordt ingedrukt, geregistreerd. Dit resulteert in een gedetailleerd overzicht van alle toetsen die een student heeft ingedrukt. De analyse van deze toetsaanslaggegevens kan inzicht geven in het schrijfproces van studenten. Er is echter een grote kloof tussen deze fijnmazige toetsaanslagen en de onderliggende cognitieve schrijfprocessen. Daarnaast is het nog grotendeels onbekend hoe deze gedetailleerde toetsaanslaggegevens kunnen worden gebruikt om docenten een bruikbaar inzicht te geven in het schrijfproces. Daarom is het doel van het huidige proefschrift om te identificeren hoe toetsaanslaganalyse kan worden ingezet om bruikbaar inzicht te verwerven in de schrijfprocessen van studenten. Dit doel is uiteengezet in vier stappen: (1) het identificeren van de behoeften van belanghebbenden, (2) het bepalen van de mogelijkheden en beperkingen van toetsaanslaganalyse, (3) het verkrijgen van inzichten uit toetsaanslaggegevens, (4) het operationaliseren van inzichten uit toetsaanslaggegevens. Deze vier stappen zijn verdeeld over zes onderzoeken die zes hoofdstukken beslaan.

Identificeren van de behoeften van belanghebbenden

In **Hoofdstuk 2** hebben we eerst de behoeften van de belanghebbenden geïdentificeerd. Hoewel het verzamelen van toetsaanslagen een groot aantal meetgegevens over het schrijven van studenten kan opleveren, is er weinig bekend over de kritische indicatoren voor het geven van geautomatiseerde feedback. Daarom beschrijft dit hoofdstuk een participatieve benadering, voor de identificatie van indicatoren van het schrijfproces die zinvol zijn voor belanghebbenden in het onderwijs om instructie en/of feedback te geven over de schrijfprocessen van studenten. In totaal werden vijf participatiesessies gehouden met vijf verschillende groepen belanghebbenden: bachelor studenten, PhD-studenten, docenten, schrijfwetenschappers en docentopleiders. De resultaten laten zien dat alle belanghebbenden indicatoren wensen uit de verschillende hoofdprocessen van het schrijven: plannen, formuleren, reviseren en zelfregulatie. Voorbeelden van de indicatoren zijn: informatie over de planningsstrategieën van studenten, informatie over hoe studenten bronnen gebruiken terwijl ze schrijven, de diepte van revisies en in hoeverre de student de taak begrijpt. Het werd belangrijk geacht om zowel generieke gedragsindicatoren (bijv. schrijfsnelheid) als abstractere cognitieve indicatoren (bijv. kritisch denken) en het verloop van gedragsindicatoren over tijd (bijv. spreiding van revisies) te extraheren. Deze bevindingen geven prioriteit aan welke indicatoren moeten worden geanalyseerd. Bovendien werd aangetoond dat de niveaus waarop deze indicatoren werden besproken, evenals de gebruikte terminologie, verschilden tussen de groepen belanghebbenden. Dit benadrukt verder de behoefte aan een mensgerichte, participatieve benadering voor het ontwerpen en ontwikkelen van computerprogramma's voor het schrijven.

Bepalen van de mogelijkheden van toetsaanslaganalyse

Na het identificeren van de behoeften van de belanghebbenden, bepaalden we wat technisch haalbaar is, gezien de beschikbare toetsaanslaggegevens. Om beter te begrijpen hoe toetsaanslaggegevens relateren aan de onderliggende cognitieve schrijfprocessen, onderzoeken we in **Hoofdstuk 3** de sensitiviteit van veelgebruikte toetsaanslagvariabelen voor taken met een verschillende cognitieve belasting. Twee datasets met toetsaanslaggegevens werden geanalyseerd: één bestaande uit een kopieertaak en een e-mailschrijftaak, en één met een groter verschil in cognitieve belasting: een kopieertaak en een academische samenvattingstaak. De verschillen tussen taken zijn gemodelleerd met Bayesiaanse lineair mixed effect

modellen. Posterieure distributies werden gebruikt om de sterkte en richting van de taakeffecten over variabelen en datasets te vergelijken. De resultaten toonden aan dat het gemiddelde van alle intertoetsaanslagintervallen stabiel was voor alle taken. Variabelen met betrekking tot de tijd tussen woorden en de tijd tussen (bij)zinnen verschilden alleen tussen de kopieertaak en de academische samenvattingstaak. Ten slotte verschilden de toetsaanslagen met betrekking tot het aantal woorden, het aantal revisies en de totale tijd van taak tot taak in beide datasets. Concluderend geven de resultaten aan dat de laatstgenoemde kenmerken gerelateerd zijn aan cognitieve belasting of taakcomplexiteit. Daarom kunnen deze gebruikt worden om inzicht te krijgen in de schrijfprocessen van studenten. Bovendien laten de bevindingen zien dat toetsaanslagen gevoelig zijn voor verschillen in schrijftaken.

Voor het geven van geautomatiseerde feedback over de schrijfprocessen van studenten, moeten we bepalen of toetsaanslagvariabelen kunnen worden gebruikt om tijdens het schrijven studenten te identificeren die moeilijkheden hebben ('risico-studenten'). Daarom verkennen we in **Hoofdstuk 4** de relatie tussen de toetsaanslaggegevens en de schrijfkwaliteit. Uit onderzoek blijkt dat toetsaanslaggegevens gebruikt kunnen worden om schrijfkwaliteit te voorspellen *nadat* een conceptversie is ingediend. In dit hoofdstuk wordt onderzocht of toetsaanslaggegevens ook gebruikt kunnen worden om de schrijfkwaliteit *tijdens* het schrijfproces te voorspellen. Toetsaanslaggegevens werden verzameld van 126 studenten (met Engels als tweede taal), die een getimede, academische samenvattingstaak uitvoerden. Het eindcijfer werd gebruikt als maat voor de schrijfkwaliteit. Op basis van de literatuur zijn 54 kenmerken van toetsaanslagen geëxtraheerd. Zelflerende (*machine learning*) modellen (regressie en classificatie) werden gebruikt om het eindcijfer te voorspellen en om risico-studenten op verschillende momenten tijdens het schrijfproces te classificeren. In tegenstelling tot eerdere studies werkten geen van de modellen beter dan de basismeting voor het voorspellen van het eindcijfer (regressie) nadat het volledige schrijfproces was voltooid. Bovendien presteerden de modellen voor het voorspellen van een risico-student (classificatie) slechts iets beter dan de basismeting. Ten slotte veranderde de relatie tussen de kenmerken van het schrijfproces en de schrijfkwaliteit over de tijd van het schrijfproces. Dit benadrukt het belang van het analyseren van de toetsaanslagen in relatie tot tijd. Bovendien toont dit aan dat de relatie tussen toetsaanslaggegevens en schrijfkwaliteit vrij beperkt is. Daarom verschuift de focus in het vervolg van het proefschrift naar het gebruik van toetsaanslaggegevens voor het modelleren van schrijf*processen* in plaats van schrijf*kwaliteit*.

Inzicht verwerven uit toetsaanslaggegevens

Met de verworven inzichten in de behoeften van de belanghebbenden en de mogelijkheden en beperkingen van de toetsaanslaggegevens gaan we vervolgens over tot het modelleren van de schrijfprocessen van studenten. We richten ons specifiek op het revisieproces, aangezien belanghebbenden aangeven dat dit één van de meest gewenste processen is om inzicht in te krijgen, dat ook direct waarneembaar is in de toetsaanslaggegevens.

In **Hoofdstuk 5** beschrijven we eerst een methodologische benadering voor het grondig bestuderen van revisies. Gezien het belang van revisie tijdens het schrijven, is revisie een veel bestudeerd thema in het schrijfonderzoek. Onderzoekers hebben verscheidene revisiemodellen ontwikkeld en er zijn verschillende taxonomieën gebruikt om revisie te meten in empirische studies. Door de huidige vooruitgang in het verzamelen en analyseren van gegevens is het mogelijk revisies nader te bestuderen. Een specifieke aanpak hiervoor ontbreekt echter nog. Daarom biedt het hoofdstuk een uitgebreide productgeoriënteerde en procesgeoriënteerde annotatietabel van revisies. De gepresenteerde annotatietabel bestaat uit tien eigenschappen van revisies: interne/externe verwerking, oorzaak, oriëntatie, evaluatie, actie, linguïstische categorie, ruimtelijke locatie, temporele locatie, duur en sequentie. Voor elk van deze eigenschappen geven we aan hoe variabelen over deze eigenschappen handmatig of automatisch kunnen worden geëxtraheerd met behulp van toetsaanslaganalyse, schermopnames, en oogbewegingdetectie. Als bewijs van het concept hebben we laten zien dat deze annotatietabel gebruikt kan worden om revisies te annoteren van studenten in verschillende academische taken. Deze annotatietabel vormt de basis voor het bestuderen en visualiseren van revisies in de volgende hoofdstukken.

Op basis van de annotatietabel gaan we in **Hoofdstuk 6** verder in op een specifiek type revisie: de revisie van typfouten. Deze revisies kunnen een grote invloed hebben op het schrijfproces en dus ook op de analyse van het schrijfproces. Enerzijds zijn dit minimale revisies en dus minder belangrijk. Het kan dus raadzaam zijn om ze te negeren voor bepaalde onderzoeksvragen. Anderzijds is het belangrijk om deze revisies te identificeren, omdat ze (ongewild) de vloeiendheid van het schrijven kunnen doorbreken en daardoor aanleiding kunnen vormen voor andere, meer omvattende revisieprocessen. Daarom beschrijft dit hoofdstuk een procesgebaseerd model om typfouten en hun revisies automatisch te identificeren. Ten eerste worden typfouten en hun revisies gekarakteriseerd op basis van temporele eigenschappen en een aantal kenmerken van de betreffende bigrammen (o.a.

frequentie). Daarna wordt een procesgebaseerd model getraind om revisies van typfouten te herkennen in toetsaanslagen van een kopieertaak. Ten slotte wordt dit model geëvalueerd in een meer normaal-functionele schrijftaak: het schrijven van een synthesetekst op basis van bronnen. De resultaten laten zien dat het mogelijk is typfouten te identificeren met alleen toetsaanslaggegevens, en dan met name in een kopieertaak. De resultaten generaliseerden echter veeleer beperkt naar de meer normaal-functionele schrijftaak. Daarom moeten we voor het modelleren van revisies van typfouten de nauwkeurigheid van de automatisch lerende-modellen verder verbeteren.

Operationaliseren van inzichten uit toetsaanslaggegevens

Tot slot, met de inzichten verkregen uit de modellen van de toetsaanslaggegevens keren we terug naar de belanghebbenden. In **Hoofdstuk 7** bepalen we hoe deze modellen kunnen worden gepresenteerd en geïntegreerd in het vakontwerp, om uiteindelijk het leren en onderwijzen van schrijven te verbeteren. Gegevens over de leerprocessen van studenten worden vaak weergegeven in zogenaamde leerdashboards. Echter, het simpelweg verstrekken van gegevens aan leerkrachten over het leerproces van studenten is niet per se gunstig voor het verbeteren van het leren; de informatie moet ook aanzetten tot actie. Onlangs is een mensgerichte aanpak voorgesteld als oplossing om effectieve en activerende dashboards te ontwerpen. Daarom gebruiken we deze mensgerichte benadering om een *interpreteerbaar* en *activerend* dashboard te ontwerpen om inzicht te geven in de revisieprocessen van studenten. Het ontwerp bestond uit drie iteratieve stappen. Eerst werden visualisaties van de revisieprocessen van studenten gemaakt op basis van toetsaanslaggegevens geëvalueerd met schrijfonderzoekers. Ten tweede werden de herziene visualisaties gebruikt om een papieren prototype van het dashboard te ontwerpen binnen een focusgroepsessie met docenten academisch schrijven. Ten slotte werd het papieren prototype omgevormd tot een digitaal prototype en door andere schrijfdocenten geëvalueerd in individuele gebruikerstests gecombineerd met interviews. De resultaten lieten zien dat deze aanpak nuttig was voor het ontwerpen van een interpreteerbaar dashboard. De docenten beschreven ook een grote verscheidenheid aan mogelijke didactische toepassingen, variërend van individuele coaching tot klassikale instructie. Zodoende opent het dashboard een nieuw perspectief om schrijfanalyses naar de klas te brengen en studenten uit te nodigen om na te denken over concrete representaties van hun schrijfprocessen.

Discussie en conclusie

Op basis van de primaire focus op toestaanslaganalyse en de verscheidenheid aan methoden die in elk van de onderzoeken werden gebruikt in dit proefschrift, zijn verschillende reflecties opgesteld. Voor het gebruik van toetsaanslaganalyse in de onderwijspraktijk is een gepersonaliseerde benadering nodig, aangezien er geen eenduidige aanpak is om het schrijfproces van studenten te verbeteren. Bovendien moet er een geschikte technische infrastructuur en omgeving worden ontworpen, die de privacy van de studenten en de transparantie van gegevens waarborgt; de lasten voor leerkrachten verlicht; en weinig datageletterheid vereist. Het gebruik van toetsaanslaganalyse in schrijfonderzoek geeft verschillende mogelijkheden voor toekomstige onderzoek. Ten eerste kunnen de datagestuurde benaderingen (zoals gebruikt in dit proefschrift) worden gebruikt voor beter schaalbare analyses (bijv. in vergelijking met handmatige annotatie). Zo kunnen ze worden gebruikt om aandachtspunten in de toetsaanslaggegevens te identificeren en licht te werpen op constructen die theoretisch moeilijk te definiëren zijn. Ten tweede—omdat toetsaanslaggegevens veranderen over tijd—is het belangrijk om ook de temporele aspecten van toetsaanslaggegevens in overweging te nemen. Tot slot, gezien de sensitiviteit van de toetsaanslaggegevens is het belangrijk om deze gegevens te combineren en specifieke zelflerende (*machine learning)* algoritmen of meer geavanceerde statistische methodes te gebruiken. Dit zal resulteren in robuustere bevindingen die gemakkelijker te relateren zijn aan de onderliggende cognitieve processen die het schrijfproces kenmerken.

In dit proefschrift toonden we aan dat het belangrijk is om zowel de technische mogelijkheden van de toetsaanslaggegevens als de behoeften van de belanghebbenden in overweging te nemen om een bruikbaar inzicht te krijgen in de schrijfprocessen van studenten. Middels een datagedreven benadering kunnen automatische en schaalbare inzichten worden verkregen in de revisieprocessen van studenten. Door vervolgens ook een mensgerichte benadering te gebruiken, kunnen deze inzichten worden omgezet in *zinvolle* inzichten met didactische toepassingen.

Acknowledgements

Just like students' writing is highly influenced by the context or their writing environment, so has been (the writing of) my dissertation. Several key persons and key events have contributed to my dissertation and/or to my views on research more generally. You have all got a special place on my desk (find yourself on the cover!).

My first desk(s)

My curiosity and ambition for conducting research really took off during my studies at Eindhoven University of Technology. From my small dorm desk, crafted by my granddad, I spent hundreds of hours studying, being inspired by all the brilliant minds around me. Two lecturers I want to name in particular, are Chris Snijders and Martijn Willemsen. Both of you taught me to work hard and to critically reflect on my own work as well as that of others. You continuously challenged me to take that one step further, driving me to achieve things I thought I would not be capable of. After my Master's, I exchanged my dorm desk for two big desks at Eindhoven University of Technology: one at the Eindhoven School of Education, and one at the Human Technology Interaction group. While working on a variety of projects, I realized this could be a lifestyle that would fit me. Here, Antoine van den Beemt showed me the ropes of a PhD life, making it an easy choice to actually start one.

My main desk

For my PhD, I moved to a desk at the department of Cognitive Science and Artificial Intelligence, overlooking the green campus of Tilburg University. This place quickly grew on me. Menno van Zaanen, as my supervisor, you introduced me to the completely new domain of natural language processing. You were always there to listen to my novel (and sometimes stubborn) ideas and ready to respond to them with a voice of reason. Even outside working hours, you created a friendly atmosphere, with the weekly Friday afternoon drinks at Tilbury's and the famous rubbed eggplants at your home barbecues. Despite your move

to South Africa, you remained closely involved. Pieter Spronck, even though I was already halfway in my PhD journey, you were still happy to take on the supervision. Your quick hands-on mentality made sure you were up to speed in no time allowing me to finish on time.

My desk was located in the middle of a big multipurpose office. Although mostly used for working, it also occasionally transformed into a local gym, a *poffertjes* bakery, a psychologist's 'couch', and (from our neighbours' perspective) a henhouse. Chrissy, sitting across me, sometimes hiding behind my screen, was always there with the best advice to spice up my English and make my emails less blunt. Too few centimeters behind me, I often bumped into Marlies. Even though you are from such a different background, you never minded listening to my ideas and were quick to find a hole in my argumentation or a solution to a problem. I am very glad you will have my back once again at my defense. Seated right of me was Thia, always happy to share her beautiful views on the world. In the corner, Marieke's enthusiastic vibe never failed to make me laugh. I so much enjoyed all your treats, souvenirs, and stories from your travels. I could not have imagined better officemates. Further down the hallway is Mirjam, always up for a good chat at the coffee machine, Tilbury's, and the occasional parties. It's a shame our schedules withheld us from sharing more train rides to and from Eindhoven. As my co-author on the carrot article, I could not think of anyone more suited to be the additional pair of eyes and ears at my defense.

In the last half year of my PhD I moved my desk one floor up to the department of Communication and Cognition at Tilburg University, where I started as a lecturer. Maria Mos, just one door away, you really guided me on how to manage teaching alongside my research. Emiel van Miltenburg, even though your recent fatherhood meant the times we *actually* shared the office were sparse, you still showed me how the interplay between research and teaching can strengthen one another.

My international desks

During my PhD, I was lucky enough to have access to some international desks as well. The Australia Awards Endeavour Research Fellowship allowed me to work for half a year at the Connected Intelligence Center, part of the University of Technology, Sydney. Here, I worked with some of the biggest names in the learning analytics community. Roberto Martinez-Maldonado, Simon Knight, and Simon Buckingham Shum, you really made me

feel welcome and taught me a lot about qualitative research and ensuring (educational) impact. You introduced me to new people at every possible moment, showing that research is even more effective and fun if you make it a team effort.

Back at Tilburg, I continued this team effort by frequenting a desk within one of the beautiful historical buildings of the University of Antwerp, the home base of the EARLI Emerging Field Group EarlyWritePro. Luuk Van Waes, from behind your small glasses and warm scarf, you added the theoretical layer to my work. I really enjoyed the long discussions and disagreements we had about Inputlog, keystrokes, and the writing process, especially when Mariëlle Leijten added yet another flavor to the discussion. Towards the end of my PhD, you shifted your guidance towards my future—discussing the two types of professor I can be, and introducing me to all the other responsibilities that are part of an academic career, including editorial tasks, supervision, and management. I am very glad this resulted in you being added to my supervisory team and my PhD being changed into a joint PhD.

One (very) temporary desk was my desk at Iowa State University. Evgeny Chukharev-Hudilainen and Emily Dux Speltz, it was a blast working with you. I am still amazed by the amount of work we got done in just two weeks, and all the new ideas we came up with.

Because of all of you, all of these offices felt like home, making the perfect environment for writing this dissertation. Thank you, and I hope to share an office with you again soon!

My dustbin

During my PhD, I liked to scribble down my ideas on paper or a whiteboard first. Most of them thankfully ended in the dustbin. Luuk and Chrissy, especially for you the bloopers; the titles that did not (yet) make it into print (and perhaps never should):

- Are you a keyboard? Because you really are my type.
- It takes two to type/typo.

My keyboard

Given the topic of my dissertation, I have been very much aware of my own writing and language in general. Given my diverse supervisory team, I got access to two both Flemish and South African and added some exotic new words to my vocabulary. The South African word for keyboard, *sleutelbord*, gave me the perfect translation for my title.

My desk's environment

Not all work happened at one of these desks. Much of my inspiration also came from random (and not so random) academics. In particular, I would like to thank: Janneke, Diana, Marije, and all the academic writing teachers I was lucky enough to interview, for your enthusiastic comments on (the teaching of) writing. Martijn, Emiel, and Marc, for showing me the bigger picture. Sander, for your random visits to our office. Drew, for your views on learning. Elske, for your dedication to teaching. Jens, for all our nights with beers and Bayesian stats.

Several others have always been there to listen to my ideas and for moral support. Marlies, Gemma, and Renée, you were never too tired to talk through my deliberations, even if we discussed these several times before. HTI guys, Vreemde tsjetroem, Sensei, and all board game fellows, you were always there to take my mind *off* of my dissertation and make some space to look at it from a distance. Bauke, as my big sister, you always understood so much more about the academic life, and it is always great to talk through all the dilemmas with you. Last but not least, mam, bedankt voor je eeuwige bereidheid om te helpen organiseren en structureren. Pap, bedankt dat je altijd met volle interesse mijn onderwerpen toehoort. Zonder jullie had ik al deze kansen nooit aangedurfd.

My home desk

The final weeks of my PhD added yet another desk to the list: the home office desk. Rob, in this period you were closer than ever to my PhD (probably even closer than you wanted to). Still, as always, you gave me all the space and support I needed to finish this dissertation. You are always there to give me that extra push to keep going, as well as that extra pull when it was time to leave my desk. You believed in me from the start and I hope I will always stay your 'professor'.

About the Author

Rianne Conijn was born on January 17, 1992 in Alkmaar, The Netherlands. She obtained her Bachelor's degree in Innovation Sciences (*cum laude)* and Masters's degree in Human Technology Interaction (*cum laude*) at Eindhoven University of Technology. During her studies she developed an interest in the analysis of students' (online) learning behavior. After her studies, she worked as a junior researcher on the analysis of students' behavior in learning management systems and MOOCs at Eindhoven School of Education and the Human Technology Interaction group at Eindhoven University of Technology.

In 2016, she started as a PhD student at the department of Cognitive Science and Artificial Intelligence at Tilburg University on the analysis of students' writing processes. In 2018, this became a joint PhD with the department of Management at the University of Antwerp, Belgium. During her PhD, she was awarded the Endeavour Research Fellowship, allowing her to conduct research for six months within the Connected Intelligence Center at the University of Technology, Sydney, Australia. In addition, she was a junior researcher in the EARLI Emerging field group EarlyWritePro, visiting researcher at Iowa State University, United States, co-teacher in the course Statistical programming in R, and thesis supervisor. In January 2020, she started as a university teacher at the department of Communication and Cognition at Tilburg University. As of June 2020, Rianne is working as a tenure track assistant professor at Eindhoven University of Technology within the Human Technology Interaction group. Her research focuses on modeling students' behavior and the implications of these (artificial intelligence) models in education, including transparency, explainability, and trust.

List of Publications

Journal papers (peer reviewed)

Knight, S., Abel, S., Shibani, A., Goh, Y. K., Conijn, R., Gibson, A., ... Buckingham Shum, S. (in press). Are you being rhetorical? An open corpus of machine annotated rhetorical moves. *Journal of Learning Analytics*.

Jansen, R. S., Van Leeuwen, A., Janssen, J., Conijn, R., & Kester, L. (2020). Supporting learners' self-regulated learning in massive open online courses. *Computers & Education*, *146*. https://doi.org/10.1016/j.compedu.2019.103771

Conijn, R., Roeser, J., & van Zaanen, M. (2019). Understanding the keystroke log: The effect of writing task on keystroke features. *Reading and Writing*, *32*(9), 2353–2374. https://doi.org/10.1007/s11145-019-09953-8

Conijn, R., van Zaanen, M., Leijten, M., & Van Waes, L. (2019). How to typo? Building a process-based model of typographic error revisions. *Journal of Writing Analytics*, *3*, 69–95.

Cook, C., Conijn, R., Antheunis, M., & Schaafsma, J. (2019). For whom the gamer trolls: empirical model of trolling in the online gaming context. *Journal of Computer-Mediated Communication*. https://doi.org/10.1093/jcmc/zmz014

Theelen, H., Willems, M., Van den Beemt, A., Conijn, R., & den Brok, P. (2019). Virtual internships in blended environments to prepare preservice teachers for the professional teaching context. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.12760

Conijn, R., Van den Beemt, A., & Cuijpers, P. (2018). Predicting student performance in a blended MOOC. *Journal of Computer Assisted Learning*, *34*(5), 615–628. https://doi.org/10.1111/jcal.12270

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Journal papers (under review)

Conijn, R., Martinez-Maldonado, R., Knight, S., Buckingham Shum, S., Van Waes, L., & van Zaanen, M. (under review). *How to provide automatic feedback on the writing process? A participatory approach to writing analytics design.*

Conijn, R., Cook, C., van Zaanen, M., & Van Waes, L. (under review). *Early prediction of writing quality using keystroke logging.*

Conijn, R., Dux Speltz, E., van Zaanen, M., Van Waes, L., & Chukharev-Hudilainen, E. (under review). *A product and process oriented tagset for revisions in writing.*

Conference contributions in proceedings (peer reviewed)

Conijn, R., Van Waes, L., & van Zaanen, M. (2020). Human-centered design of a dashboard on students' revisions during writing. In *Conference proceedings of the 14th European Conference on Technology Enhanced Learning, EC-TEL* (pp. 1–15). https://doi.org/10.1007/978-3-030-57717-9_3

Conijn, R., Dux Speltz, E., van Zaanen, M., Van Waes, L., & Chukharev-Hudilainen, E. (2020). A process-oriented dataset of revisions during writing. In *Proceedings of the 12th language resources and evaluation conference* (pp. 356–361). Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.44

Conijn, R., van Zaanen, M., & Van Waes, L. (2019). Don't wait until it is too late: The effect of timing of automated feedback on revision in esl writing. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Learning with Meaningful Technologies, Conference proceedings of the 14th European Conference on Technology Enhanced Learning, EC-TEL* (pp. 577–581). Delft, The Netherlands. https://doi.org/10.1007/978-3-030-29736-7_43

Conijn, R., Van der Loo, J., & van Zaanen, M. (2018). What's (not) in a keystroke? Automatic discovery of students' writing processes using keystroke logging. In *Companion Proceedings of the 8th International Conference on Learning Analytics & Knowledge.* Sydney, Australia.

Conijn, R., & van Zaanen, M. (2017a). Identifying writing tasks using sequences of keystrokes. In *Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning* (pp. 28–35). Eindhoven, The Netherlands.

Conijn, R., & van Zaanen, M. (2017b). Trends in student behavior in online courses. In *Proceedings of the 3rd International Conference on Higher Education Advances* (pp. 649–657). Valencia, Spain. https://doi.org/10.4995/HEAd17.2017.5337

Book chapters (peer reviewed)

Conijn, R., Nij Bijvank, W., Snijders, C., Kleingeld, A., & Matzat, U. (2018). From raw to ready-made data. a hands-on manual for pre-processing Learning Management System log data for learning analytics. In C. Stuetzer, M. Welker, & M. Egger (Eds.), *Computational Social Science in the Age of Big Data: Concepts, Methodologies, Tools, and Applications.* Cologne, Germany: Herbert von Halem Verlag.

SIKS Dissertation Series

2016

1. Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
2. Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
3. Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
4. Laurens Rietveld (VU), Publishing and Consuming Linked Data
5. Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
6. Michel Wilson (TUD), Robust scheduling in an uncertain environment
7. Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
8. Matje van de Camp (UVT), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
9. Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10. George Karafotias (VU), Parameter Control for Evolutionary Algorithms
11. Anne Schuth (UVA), Search Engines that Learn from Their Users
12. Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13. Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14. Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15. Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16. Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17. Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18. Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19. Julia Efremova (TUE), Mining Social Structures from Genealogical Data
20. Daan Odijk (UVA), Context & Semantics in News & Web Search
21. Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22. Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23. Fei Cai (UVA), Query Auto Completion in Information Retrieval
24. Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25. Julia Kiseleva (TUE), Using Contextual Information to Understand Searching and Browsing Behavior
26. Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27. Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28. Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29. Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30. Ruud Mattheij (UVT), The Eyes Have It
31. Mohammad Khelghati (UT), Deep web content monitoring
32. Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations

33. Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34. Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35. Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36. Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37. Giovanni Sileno (UVA), Aligning Law and Action - a conceptual and computational inquiry
38. Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39. Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40. Christian Detweiler (TUD), Accounting for Values in Design
41. Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42. Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43. Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44. Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45. Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46. Jorge Gallego Perez (UT), Robots to Make you Happy
47. Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48. Tanja Buttler (TUD), Collecting Lessons Learned
49. Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50. Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

1. Jan-Jaap Oerlemans (UL), Investigating Cybercrime
2. Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
3. Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
4. Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
5. Mahdieh Shadi (UVA), Collaboration Behavior
6. Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
7. Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
8. Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
9. Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10. Robby van Delden (UT), (Steering) Interactive Play Behavior
11. Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12. Sander Leemans (TUE), Robust Process Mining with Guarantees
13. Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14. Shoshannah Tekofsky (UVT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15. Peter Berck (RUN), Memory-Based Text Correction
16. Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17. Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18. Ridho Reinanda (UVA), Entity Associations for Search
19. Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20. Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21. Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22. Sara Magliacane (VU), Logics for causal inference under uncertainty
23. David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24. Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25. Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26. Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27. Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28. John Klein (VU), Architecture Practices for Complex Contexts
29. Adel Alhuraibi (UVT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT
30. Wilma Latuny (UVT), The Power of Facial Expressions
31. Ben Ruijl (UL), Advances in computational methods for QFT calculations
32. Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33. Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34. Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35. Martine de Vos (VU), Interpreting natural science spreadsheets
36. Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37. Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38. Alex Kayal (TUD), Normative Social Applications
39. Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40. Altaf Hussain Abro (VU), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41. Adnan Manzoor (VU), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42. Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43. Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44. Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45. Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46. Jan Schneider (OU), Sensor-based Learning Support
47. Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48. Angel Suarez (OU), Collaborative inquiry-based learning

2018

1. Han van der Aa (VU), Comparing and Aligning Process Representations
2. Felix Mannhardt (TUE), Multi-perspective Process Mining
3. Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
4. Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
5. Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
6. Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
7. Jieting Luo (UU), A formal account of opportunism in multi-agent systems
8. Rick Smetsers (RUN), Advances in Model Learning for Software Systems
9. Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10. Julienka Mollee (VU), Moving forward: supporting physical activity behavior change through intelligent technology

11. Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12. Xixi Lu (TUE), Using behavioral context in process mining
13. Seyed Amin Tabatabaei (VU), Computing a Sustainable Future
14. Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15. Naser Davarzani (UM), Biomarker discovery in heart failure
16. Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17. Jianpeng Zhang (TUE), On Graph Sample Clustering
18. Henriette Nakad (UL), De Notaris en Private Rechtspraak
19. Minh Duc Pham (VU), Emergent relational schemas for RDF
20. Manxia Liu (RUN), Time and Bayesian Networks
21. Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22. Eric Fernandes de Mello Araujo (VU), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23. Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24. Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25. Riste Gligorov (VU), Serious Games in Audio-Visual Collections
26. Roelof de Vries (UT), Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27. Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
28. Yu Gu (UVT), Emotion Recognition from Mandarin Speech
29. Wouter Beek (VU), The "K" in semantic web stands for "knowledge": scaling semantics to the web

2019

1. Rob van Eijk (UL), Comparing and Aligning Process Representations
2. Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
3. Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
4. Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
5. Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
6. Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
7. Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
8. Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
9. Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
10. Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11. Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12. Jacqueline Heinerman (VU), Better Together
13. Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14. Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15. Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16. Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17. Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18. Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19. Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20. Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21. Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22. Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture
23. Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24. Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25. Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26. Prince Singh (UT), An Integration Platform for Synchromodal Transport
27. Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses
28. Esther Kuinderman (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29. Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30. Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
31. Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
32. Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks
33. Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
34. Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming
35. Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills
36. Jian Fang (TUD), Database Acceleration on FPGAs
37. Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020

1. Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
2. Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
3. Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
4. Maarten van Gompel (RUN), Context as Linguistic Bridges
5. Yulong Pei (TUE), On local and global structure mining
6. Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
7. Wim van der Vegt (OUN), Towards a software architecture for reusable game components
8. Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
9. Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10. Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
11. Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12. Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13. Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14. Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15. Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16. Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17. Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18. Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19. Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20. Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21. Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
22. Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar