What's (not) in a Keystroke? Automatic Discovery of Students' Writing Processes Using Keystroke Logging

Rianne Conijn[1], Janneke van der Loo[2], Menno van Zaanen[1]
[1]Department of Cognitive Science and Artificial Intelligence
[2]Department of Cognition and Communication
Tilburg University
{m.a.conijn, j.m.vdrloo, m.m.vanzaanen}@tilburguniversity.edu

ABSTRACT: Teachers typically do not have access to students' writing processes, such as planning and revision, but only to final products. Students' writing processes can be analyzed by labor-intensive methods such as thinking aloud or by manually labelling behavior logs. This paper describes an approach to automatically extract writing processes from keystroke data. Keystroke data from 70 students writing an academic synthesis task are analyzed. A heuristic-based method is used to extract the planning and revision processes. In addition, Bayesian correlational analysis and *t*-tests are used to identify the relation between the extracted processes and students' self-reported writing style. The results show that the heuristic-based method can extract planning and revision features from keystrokes. However, no relation between the planning features and self-reported planning style and a limited relation between revision features and self-reported revision is found. Some anecdotal evidence is found that high revisers typed more revision characters than low revisors. To arrive at the fully automatic analysis of students' writing processes, future work should extract more keystroke features and evaluate their relation with the actual writing processes.

Keywords: Writing analytics, writing processes, writing strategies, keystroke analysis.

1 INTRODUCTION

Writing teachers often only have access to the final writing products constructed by the students, which does not include information about the actual writing processes. Two writing processes or strategies often used in writing research are planning and revision (Flower & Hayes, 1980). To improve writing instruction, it would be useful to have insight into these writing processes as well. Traditionally, students' writing processes were analyzed using thinking-aloud methods, self-report questionnaires, and retrospective interviews. Nowadays, with learning and writing becoming more digitalized, data about students' writing processes can be collected automatically. Keystroke logging is one tool which can be used to automatically collect students' typing behavior.

Keystroke logging in writing research has been used for a wide variety of aims. For example, keystroke logging has been used to predict essay score (Zhang, Hao, Li, & Deane, 2016), distinguish skilled versus less-skilled writers (Xu & Ding, 2014), to determine boredom and engagement (Allen et al., 2016), to assess mental ability (Van Waes, Leijten, Mariën, & Engelborghs, 2017), and to determine the tasks' cognitive load (Wallot & Grabowski, 2013). Yet, it is still considered difficult to extract higher-level writing processes from keystroke logs (Baaijen, Galbraith, & De Glopper, 2012; Leijten & Van Waes, 2013).

Some researchers tried to relate keystrokes to higher-level writing processes. Van Waes, Van Weijnen, and Leijten (2014) analyzed the relation between keystrokes and students' self-reported learning style when writing a bad news letter. No relation was found between the keystroke features (pauses between keys, characters produced) and learning style. Baaijen and colleagues (2012) did find a relation between keystroke features (timing of pauses, timing and place of revisions) and type of revisions. Using principal components analysis, five main components were derived: planned sentence production, within-sentence revision, revision of global structure, and (tentatively labeled) post draft revision and careful word choice. Lastly, Tillema, Van den Bergh, Rijlaarsdam, and Sanders (2011) related keystrokes to planning and revision behavior, with manual labels. In addition, they compared this behavior with self-reported planning and revision styles. High planners were found to read less often, were more likely to plan at the start than in the end, produced more, and revised more, compared to low planners. High revisers were found to read their own text less often, compared to low revisers.

In contrast to the studies above, we will automatically extract both planning and revision processes from keystroke data obtained during an academic synthesis task using a heuristic-based method. In addition, we try to relate the extracted writing processes to students' self-reported writing style.

2 METHOD

2.1 Participants

In this study, first year undergraduate communication and information sciences students from Tilburg University, who followed the course Academic Dutch were asked to complete an academic synthesis task in Microsoft Word. Demographics and self-reported writing styles were collected in the form of a pre-test. In total, 74 participants provided informed consent and participated in this study. The academic synthesis task is a mandatory task in the course. The task aims to practice writing an academic introduction. The participants were asked to read three short academic texts at home first. Thereafter, in the classroom, the participants got 30 minutes to write (the start of) an introduction in Dutch (their native language) based on these three academic texts. They were asked to type everything and to not make written notes. During this task, keystrokes were collected. After the task, the students were allowed to finish the task at home, before handing it in.

2.2 Writing style self-report

Students' self-reported writing style were collected with the Writing Style Questionnaire (Kieft et al., 2006; 2008). This questionnaire consists of 13 statements on planning, 12 statements on revision, and 12 filler statements. All questions were answered on a five-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). This questionnaire provides two scores, a score on planning and revising style. Participants could score equal on both styles, or one of the styles could be dominant. The internal consistency was similar to that found by Kieft et al. (2006; 2008), with a Cronbach's alpha of .73 for the planning dimension and .69 for the revision dimension. The planning and revision scores were only moderately correlated (*r* = .39), indicating that the scores can be analyzed separately. The participants scored somewhat higher on revision (*M* = 3.5, *S.D.* = 0.48) compared to planning (*M* = 3.0, *S.D.* = 0.51). Median split was used to recode the planning and revision scores into binary variables, to analyze the differences in planning and revision between high and low planners and high and low revisers.

2.3 Keystroke data feature extraction

The keystrokes were collected with Inputlog (Leijten & Van Waes, 2013), which logs every key pressed and the times of the key press and key release. On average, the participants pressed 2967 keys (*S.D.* = 1076), which resulted in 2567 characters (*S.D.* = 1178) produced. The final document (after the 30 minutes) consisted on average of 1809 characters (*S.D.* = 895), indicating that a fair amount of revision took place. Planning and revision features were extracted from the keystrokes using a heuristic-based method. Here, rules are used to denote parts of the overall keystroke sequences as either 'planning' or 'revision'. On average, 2.9% of all the keystrokes were labeled as planning, and 16.4% of the keystrokes as revision.

The rules for labeling a sequence as *planning* included: the first phrases or non-complete sentences (sentences without a period) with at least 20 characters. Thus, typing a heading such as "inleiding opdracht 1" (*Dutch for: Introduction assignment 1*), would not be considered planning. Note, this only includes initial planning, not the planning in the middle of the writing processes when already some full sentences are produced. Based on these rules, four planning features were extracted: initial pause time (time until the first keystroke), number of plan characters, plan time, and plan character ratio (number of characters planned/total number of characters).

Rules for labeling a sequence as *revision* included all consecutive keystrokes where the next keystroke resulted in a lower document length, i.e., something was removed. Based on these rules, four revision features were extracted, which were similar to the planning features: the number of revisions, number of characters revised, revision time, and revision character ratio (number of characters revised/total number of characters).

Data from four participants were removed. One participant wrote in English instead of Dutch. In addition, three outliers (features more than three *S.D.* above the mean) were removed because these had a significant influence on the results. In total, data from 70 participants were left for analysis.

2.4 Analysis

The relation between the extracted planning and revision features with the self-reported writing style was analyzed using Pearson's correlation analysis, and evaluated with Bayes Factor, calculated in R with a Jeffreys-Zellner-Siow (JZS) prior set-up (Wetzels & Wagenmakers, 2012). The Bayes Factor (BF_{10}) quantifies the evidence in favor of one hypothesis, over an alternative hypothesis. The number indicates how much more (un)likely the data are to have occurred under the alternative hypothesis, compared to the null hypothesis. Next to correlational analysis, we analyzed whether high self-reported planners showed significantly more planning than low planners, and whether high self-reported revisers showed significantly more revisions compared to low revisers. Bayesian *t*-tests, implemented using the BEST package in R (Kruschke & Meredith, 2017), were used to compare the planning and revision features between the low/high planners and low/high revisers, respectively.

3 RESULTS

The correlational analysis showed that there is moderate evidence against a correlation between planning score and initial pause time and planning (r = -.09, BF_{10} = 0.13), the number of plan characters (r = .13, BF_{10} = 0.16), plan time (r = .07, BF_{10} = 0.10), and plan character ratio (r = .34, BF_{10} = 6.14). Here, a Bayes Factor of 0.13 indicates that the data are 1 / 0.13 = 7.7 times more likely to have occurred under the null hypothesis H_0 (no correlation) than under the alternative hypothesis H_1 (correlation). Thus, none of the planning features seem to be correlated with the self-reported planning score. For revision, a moderate evidence against a correlation was found between revision score and the number of revisions (r = .06, BF_{10} = 0.10), number of revision characters (r = .11, BF_{10} = 0.14), revision time (r = .08, BF_{10} = 0.12), and revision character ratio (r = .07, BF_{10} = 0.11).

In addition, Bayesian *t*-tests were conducted to analyze whether high self-reported planners or high revisers indeed showed more planning or revision compared to low planners or low revisers. No significant differences were found between the planning features for high/low planners (Table 1). All 95% highest density interval (HDI) included zero, thus the differences between the two means were not significantly different from zero. The low Bayesian Factors also support the evidence for the null model (no differences between the means). Likewise, all 95% HDIs for the reviser features included zero, indicating no significant differences between the revision features for high/low revisers (Table 2). However, the Bayes Factor does show some anecdotal evidence for a difference in revision characters. Further inspection indeed showed that 92% of the HDI was above zero, thus there is a 92% probability that the mean of the number of revision characters is higher for high revisers, compared to low revisers.

Table 1: Bayesian *t*-tests planning features

Feature	Overall M (S.D.)	High Planner M (S.D.)	Low planner M (S.D.)	95% HDI	BF_{10}
Initial pause time (s)	305 (105)	290 (87)	318 (118)	[-70.9, +25.4]	0.41
Number of plan characters	74 (116)	82 (119)	66 (115)	[-13.3, +17.3]	0.28
Plan time (s)	88 (151)	83 (156)	93 (149)	[-4.84, +11.4]	0.26
Plan character ratio (%)	3.5% (7.3%)	3.8% (6.7%)	3.2% (8.0%)	[-0.3%, +1.4%]	0.26

Table 2: Bayesian *t*-tests revision features

Feature	Overall M (S.D.)	High reviser M (S.D.)	Low reviser M (S.D.)	95% HDI	BF_{10}
Number of revisions	117 (55)	127 (67)	106 (44)	[-12.1, +46.0]	0.49
Number of revision characters	817 (653)	993 (873)	677 (358)	[-82.8, +460]	1.47
Revision time (s)	146 (64)	153 (69)	140 (59)	[-16.8, +42.4]	0.32
Revision character ratio (%)	30% (14%)	33% (15%)	28% (12%)	[-2.0%, +12%]	0.69

4 CONCLUSION

The current work described an approach to extract planning and revision processes from keystroke logs using a heuristic-based method. We showed that revision and planning processes can at least to some extent be extracted from keystrokes. In addition, properties from these processes were related to the self-reported planning and revision writing styles. In future work, we will explore ways to extract more (detailed) processes from keystrokes. For example, we will include planning in the middle of the writing task (we now only included initial planning) or different types of revision, such as surface and meaning revisions (Faigley & Witte, 1981). Yet, these features might be harder to accurately identify using a heuristic-based method.

The extracted keystroke features showed limited to no relation with the self-reported writing style. Only some evidence is found that high revisers use more revision characters, compared to low revisers. These findings are consisted with the findings Van Waes and colleagues (2014), who did not find a relation between self-reported learning style and keystroke features. However, Tillema et al. (2011) found a relation between self-reported writing style and keystroke features. Yet, in their study, the keystrokes were manually labeled with writing processes. This indicates that there is some relation between the self-reported writing style and actual planning and revision behavior. This would suggest that we are not yet extracting the right features from the keystrokes which represent planning and revision behavior. To evaluate whether the extracted features indeed relate to writing and revision processes, in future work, we will manually code the dataset to evaluate the extracted features. In addition, this labeled dataset can be used to automatically classify a given sequence of keystrokes. This paper showed the first steps towards the automatic discovery of writing processes using keystroke logging.

REFERENCES

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). ACM.

Baaijen, V. M., Galbraith, D., & de Glopper, K. (2012). Keystroke Analysis: Reflections on Procedures and Measures. *Written Communication*, *29*(3), 246–277.

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

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

Kieft, M., Rijlaarsdam, G., & van den Bergh, H. (2006). Writing as a learning tool: Testing the role of students' writing strategies. *European Journal of Psychology of Education*, *21*(1), 17–34.

Kieft, M., Rijlaarsdam, G., & van den Bergh, H. (2008). An aptitude–treatment interaction approach to writing-to-learn. *Learning and Instruction*, *18*(4), 379–390.

Kruschke, J. K. & Meredith, M. (2017). BEST: Bayesian Estimation Supersedes the t-Test. R package version 0.5.0. https://CRAN.R-project.org/package=BEST

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.

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.

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.

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.

Wallot, S., & Grabowski, J. (2013). Typewriting Dynamics: What Distinguishes Simple From Complex Writing Tasks? *Ecological Psychology*, *25*(3), 267–280.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064.

Xu, C., & Ding, Y. (2014). An exploratory study of pauses in computer-assisted EFL writing. *Language Learning & Technology*, *18*(3), 80–96.

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). Cham: Springer International Publishing.