UMEÅ UNIVERSITET

Exploring Cognitive Processes in AI-Assisted Academic L2 Writing

Andreas Sellstone

Master's Thesis, 15 ECTS
Master's Programme (two year) in Cognitive Science, 120 ECTS
Spring 2024
Supervisors: Kirk Sullivan, Linda Sandström

Abstract

This exploratory study investigates the role of academic writing in a second language (L2) with automated writing evaluation (AWE) tools utilizing generative artificial intelligence (AI), as well as the role of such AI-assistance in cognitive writing processes. The research questions asked if there was a relationship between AWE suggestions and keystroke dynamics, how L2 writers interpreted AWE suggestions and made their decisions, and how AI-assisted writing functions within a cognitive model of writing. Six academics wrote a text in English as L2 using a keystroke logging program. The text was later revised using the AWE program InstaText while their interactions were screen recorded. The recording was then promptly followed by stimulated recall interviews. The keystroke logging data that indicates cognitive demand (e.g., pause length, pause frequency, and deletion frequency) were compared to the AWE suggestion frequency. A thematic analysis was conducted on the interviews to investigate the participants' metacognitive reflections about their decision-making process. A cognitive model of the writing process was used to map the cognitive processes with the keystroke logging data, AWE interactions, and metacognitive reflections. The findings indicate that longer pauses and fewer typos led to fewer AWE suggestions. Six types of metacognitive reflections were identified in the qualitative data, which was then interpreted together with the quantitative data through a cognitive model of the writing process.

Keywords: AI-assisted writing, artificial intelligence, keystroke logging, cognitive processes, writing processes

Sammanfattning

Denna explorativa studie undersöker akademiskt skrivande på ett andra språk (L2) med automatiserade skrivbedömningsverktyg (AWE) som använder generativ artificiell intelligens (AI), samt AI-assistansens roll i kognitiva skrivprocesserna. Studien undersökte om det fanns en relation mellan AWE-förslag och tangentloggningsdata, hur L2-skribenter tolkade AWE-förslag och fattade sina beslut, och hur AI-assisterat skrivande fungerar inom en kognitiv modell av skrivande. Sex akademiker skrev en text på engelska som L2 med ett tangentloggningsprogram. Texten reviderades sedan med AWE-programmet InstaText medan deltagarnas interaktioner spelades in på skärmen. Inspelningen följdes omedelbart av stimulated recall-intervjuer. Tangentloggningsdata som indikerar kognitiv belastning (pauslängd, pausfrekvens och raderingsfrekvens) jämfördes med frekvensen av AWE-förslag. En tematisk analys utfördes på intervjuerna för att undersöka deltagarnas metakognitiva reflektioner kring deras beslutsfattande. En kognitiv modell av skrivprocessen användes för att kartlägga de kognitiva processerna med tangentloggningsdata, AWE-interaktionerna och metakognitiva reflektionerna. Resultaten indikerar att längre pauser och färre stavningsfel ledde till färre AWE-förslag. Sex typer av metakognitiva reflektioner identifierades i de kvalitativa data, vilket tolkades tillsammans med de kvantitativa data genom en kognitiv modell av skrivande.

Nyckelord: AI-assisterat skrivande, artificiell intelligens, tangentloggning, kognitiva processer, skrivprocesser

Exploring Cognitive Processes in AI-Assisted Academic L2 Writing

New developments in generative Artificial Intelligence (AI) are reshaping our society in many ways and will continue to do so in the foreseeable future. AI models have shown themselves capable of producing highly sophisticated content including images, music, and text. This has caused worries regarding the use of AI in academia and discussion about either integrating this technology into the education system or preventing students from using such tools all together. One thing is certain, AI is increasingly being used in higher education by both students and faculty.

There have long been technological tools aimed at improving writing. Automated Writing Evaluation (AWE) systems capable of not only correcting the grammar of written text, but also improving the style and content of writing have existed for some time, one of the more popular being Grammarly. In recent years some AWEs have incorporated AI to improve AWE accuracy and give individualized and context dependent feedback (Ding & Zou, 2024). Such AI powered AWE tools are gaining popularity in academia to improve the writing of students and faculty, particularly in second language (L2) academic writing. One challenge facing AWE users is assessing the generated feedback itself to see whether a suggested improvement is stylistically appropriate to the writing context, or if the suggestions preserve the intended meaning. Some researchers have only recommended the use of AWE tools for relatively advanced English writers because of the competence required for such assessments (Koltovskaia, 2020).

The role of AI in L2 writing and cognitive models of the writing process remains obscure due to its recent developments and extended use. Consequently, the goal of this exploratory study is to investigate the role of AI powered AWE systems in cognitive models of the writing process, as well as how L2 English speakers understand the suggestions generated by the AWE system. The keystroke logging program GenoGraphiX-LOG (Caporossi et al., 2023) was used to probe the cognitive writing processes without the use of AWE, after which participants transferred their text to the AI powered AWE system InstaText (InstaText, n.d.-c). Participants used InstaText to revise their texts by rejecting, accepting, or substituting the suggestions generated by the AWE. The interaction with the AWE was screen recorded for a stimulated recall interview immediately afterward, where participants were asked to explain their thought process and reasoning for their decisions. To investigate the role of AWE in writing, Hayes and Berninger's (2014) cognitive model of writing was used to interpret the data according to specific cognitive processes and resources by mapping the model components to the observations obtained from the participants. The keystroke logging data was compared to the suggestions (RQ1-3) that the AWE proposed.

The research questions investigated in this thesis are as follows:

(1) is there an identifiable relationship between the markers of demand (pauses and deletions) identified by keystroke logging data and the improvements suggested by AWE;
(2) how do L2 writers of different academic status understand AWE suggestions, and what are the reasons for accepting, rejecting, or substituting AWE suggestions;
(3) how do AI powered AWE systems fit into the cognitive processes of L2 writing as outlined in Hayes and Berninger's model?

I will begin by providing the background of the thesis by going through current discussions and topics about the application of AI as a writing tool in academia and AWE systems, as well as keystroke logging and its use in research on the cognitive processes involved in writing, giving an overview of Hayes and Berninger's model of writing.

Technology in Academic Writing

Generative Artificial Intelligence

The "generative" in AI refers to the model's capability of producing new content, whether it be text, images, music, or even video (Feuerriegel et al., 2024). Such models have had a developmental explosion in their use and capabilities, sometimes being difficult to distinguish from human created content. Countless discussions of the use of AI in diverse contexts have surfaced because of this development. Particularly relevant to this thesis is the discussion about AI in school and academia.

More and more people are using AI to aid their writing in various respects, whether it be as a tool to help generate ideas for the structure and content of papers, or to give feedback on writing. It can even be used to generate fully fledged papers with minimal input from the user. Herbold et al. (2023) compared the quality of essays generated by the AI models ChatGPT and GPT-4 to essays written by high school students. They found that teachers rated the AI essays higher in terms of quality than the student essays. Their results revealed that, on average, ChatGPT and GPT-4 were scored 0.67 and 0.99 points higher, respectively, on a seven-point Likert scale than human writers when considering many different criteria of quality such as topic and completeness, logic and composition, and complexity. Herbold et al. subsequently argue for the integration of AI into education by adapting teaching methods, rethinking homework, and developing new educational concepts, in the same vein as what happened in mathematics after the invention of calculators. AI could then be utilized to free up time for other learning objectives.

In the ongoing and lively debate, some argue for integration and some for the exclusion of AI in schools. The former side emphasize the possibility of enhancing learning opportunities, including personalized learning experiences, and support for teachers, while the latter stress the possibility of increased plagiarism, decreased critical thinking through overreliance, as well as the perpetuation of discrimination through biased training data (Mhlanga, 2023). However, it seems that the use of AI is increasing rapidly, even among faculty. In a survey by Shaw et al. (2023), it was reported that 9% of faculty and 27% of students used AI in the spring of 2023. This rose to 22% of faculty and 49% of students in fall of the same year.

Automated Writing Evaluation Systems

Writers have long used software to automatically provide feedback on their texts, especially in terms of grammar and spelling. These tools are called automated writing evaluation (AWE) systems, and such systems assess writing by automated written corrective feedback (AWCF) or automated scoring (Ding & Zou, 2023). Writing tools such as Grammarly are of the former type, it gives feedback on written text and suggests improvements based on spelling and grammar. The latter type of AWE analyzes texts and can score them based on their readability, quality, language use, or other metrics, mostly used by teachers to aid grade written content. One such AWE that will be used in the current study is the text analyzer Text Inspector (Text Inspector, n.d.-b), which measures text properties such as lexical diversity, readability, and language quality, among others.

Research into the effectiveness of AWE systems on writing proficiency have shown positive results in EFL (English as a foreign language). A systematic analysis by Ding and Zou (2023) of three AWE systems (Grammarly, Pigai, and Criterion) stated that most of the included studies (11 out of 15) reported positive results on writing development. Thus, not only does the quality of the text where AWE is used improve, users seem to improve their writing skills in the long term as well. According to a study by Guo et al. (2022), students with lower writing proficiency tend to be more likely to accept suggestions compared to students with

higher writing proficiency. This is thought to be the result of increased skepticism toward AWCF. However, on an individual basis, there is still risk of overreliance on AWE. In a case study by Koltovskaia (2020) of two EFL students, one advanced L2 writer and one low-intermediate L2 writer, the low-intermediate participant tended to accept the AWCF whenever they were unsure about the feedback's accuracy. When users of low writing proficiency are using AWE, there seems to be a danger of both overreliance and unsubstantiated skepticism. Therefore, some researchers recommend them for relatively advanced writers.

AWE technology has utilized AI to increase the quality of AWCF in recent years, basing feedback on contextual language use such as formality and clarity, sometimes rewriting full sentences. In other words, one way AWE technology that leverages AI differs from other forms of AWE is their capacity to generate new content based on the context of the original text. One example of an AI powered AWE system is InstaText (InstaText, n.d.-c). InstaText is a program that specifically targets L2 English users and claims to be a writing assistant that "rewrites your texts like a native speaker" (InstaText, n.d.-b, para. 6). AWE without AI might suggest changes related to punctuation, grammar and sometimes offer alternative words, while AWE powered by generative AI can offer suggestions that completely rewrites the original text. InstaText is very simplified, requiring a limited set of interactions. InstaText allows the user to write or paste their text in a window to the left and generate suggestions for revisions in a window to the right. The suggestions can be accepted or rejected by the user (or substituted by one's own revisions); no other information—whether the suggestions relate to grammar, spelling, clarity, or formality—about the suggestions is given. The writing tool has been well received by some and a study by Beikian and Ganji (2022) found that 75 Persian freelance translators had statistically significant improvements in the quality of their texts when using InstaText.

InstaText was deemed suitable for the purposes of the current study because the tool uses AI to power the AWCF and specifically targets L2 and academic writers with the prospect of writing like a native speaker (InstaText, n.b.-a; InstaText, n.d.-b). Furthermore, its simple interface and usability, as indicated by user satisfaction in previous research (Beikian & Ganji, 2022), lends itself to methodological integration. In comparison to InstaText, Grammarly's AI feature (Grammarly, n.d.) does not seem to specifically target L2 writers, nor academic writers, and its interface differs significantly from other more traditional AWCF technology. For example, in Grammarly's AI feature, the user selects the originally produced text (in full or in part) and then writes instructions on how to revise the text to the AI in a popup window. The generated suggestion for the whole selected text is then evaluated and decided upon by the user. In comparison, InstaText generates multiple suggestions that each can be accepted or rejected by the user. Grammarly's feature offers a lot of freedom for the user since one can essentially ask the AI to revise the text in any way the user sees fit. This would, however, complicate the methodology of the current thesis, and the straightforward interaction that InstaText offers was considered apt the purposes of the current study.

As previously mentioned, a challenge with the use of AWE tools, especially for L2 writing, is the assessment of the generated feedback itself. There might be an increased risk of users not understanding the implications of certain suggestions since AWE programs such as InstaText specifically target L2 English writers. Therefore, one of the objectives of this study was to investigate how L2 English writing participants understand InstaText's AWCF, as well as why they accept, reject, or substitute suggestions.

Cognitive Models of Writing

There seems to be a research gap when it comes to how generative AI technologies affect cognitive writing process. With the developments of increasingly sophisticated AWE

tools and AI systems capable of functioning as writing assistants, the process of writing has the potential to transform from a decreasingly individual and autonomous process to an increasingly collaborative process between human and technology, where certain cognitive processes can be delegated to machines. This is not to say that the writing process without AI or other writing technologies is not collaborative, involving many different actors and collaborators, but rather that, with AI, novel cognitive partnerships with technology are introduced, where tasks traditionally reserved for the human mind are executed by a computer. Therefore, one of the goals of this thesis is to provide a preliminary glance into the role of AI powered AWE technology in the cognitive writing process.

A prevalent account of writing as a cognitive process is Hayes and Berninger's (2014) model. This model is composed of three levels—the control level, process level, and resource level—that systematically organize cognitive resources and processes that are employed during writing. The highest level, the control level, consists of regulatory factors that direct the cognitive processes at the process level. In the control level, the initiator is the agent that initiates the writing, whether it be a teacher or a manager giving the writer a task, or the writer themselves. The planner is whomever sets the goals of the writing task, for example, what topic to write about, and the tone and impact the writing should have on audiences. The control level also includes the writing schemas. Writing schemas are the writer's own beliefs about the properties that the text should have, what is called genre knowledge, as well as the writer's strategies on how to produce a text with those properties, what is called strategic knowledge. The strategies determine how the writing processes operate and how one interacts with the task environment.

The process level includes both cognitive writing processes and the task environment. The cognitive processes are the proposer, the translator, the transcriber, and the evaluator. The proposer suggests non-verbal ideas to be included in the text. The translator takes the non-verbal ideas and translates them into grammatical and semantic strings. The transcriber turns the strings into written text and the evaluator examines the outputs of the previous processes, making it possible to interrupt the writing process at any stage. The second part of the process level, the task environment, includes the immediate social and physical environment that affects the writing process. The social environment includes input from social sources such as teachers, critics, collaborators, and the physical environment consists of the transcribing technology, the task material, as well as the text written so far.

The last level in Hayes and Berninger's (2014) model consists of the general cognitive resources that writers use when writing: attention, since one must maintain focus on the task; long-term memory, in terms of knowledge of language, facts, and previous writing experience; working memory, to temporarily store task-related information; and reading, used when writers read their output, as well as external sources.

With Hayes and Berninger's (2014) cognitive model of writing outlined, where might AI technology come into play? AI seems to fit nicely within the task environment, however, exactly how AI fits into the model components is unclear. With the sophisticated tools that are being developed, it seems that AI can perform many of the processes engaged in writing, interacting with almost all the above outlined components. In this study the focus will be on how the AI powered AWE program InstaText fits into the components of the cognitive model of writing by Hayes and Berninger.

Keystroke Logging

Keystroke logging is a popular tool in writing research and has been used to probe cognitive writing processes (Galbraith & Baaijen, 2019). Keystroke logging provides detailed information about the writing process, primarily composed of each keystroke, pauses (the

duration between keystrokes), revisions, and text production rate (e.g., words/characters per minute). This data can be mapped to components of Hayes and Hayes and Berninger's (2014) model. For example, pauses can be indicators of planning by temporarily ceasing writing to think about what to write next or evaluating their text, which could therefore include both the proposer, the evaluator, and the translator; writing bursts (sequences of keystrokes between interruptions such as two second pauses or revisions) are linked to the translator, where thoughts are converted into language; and revisions that suggest previous employment of the evaluator that has demanded a revision by the other processes (Galbraith & Baaijen, 2019).

Researchers have inferred the depth of cognitive processing by analysing keystroke markers such as the fluency and flow of writing (Leijten & Van Waes, 2013). It might be the case that indicators such as length and number of pauses and deletions that can indicate struggle in the writing process correspond to the AI suggestions. Hence, one of the aims of this thesis is to investigate the relationship between the keystroke logging data of the writing without the AWE and the AWCF. Furthermore, the data from the interaction with the AWE program, including stimulated recall, will be analyzed and related to the components of Hayes and Berninger's (2014) model of writing, providing insight into the writing strategies employed during real-time interactions with the AWE program—revealing processes such as planning and translation of ideas, through metacognitive accounts of the writing process.

Method

Participants

Six L2 speakers of English participated in this study. The participants were grouped according to academic level: group 1 consisted of two bachelor's students, group 2 consisted of two master's students, and group 3 consisted of two academics above master's level. Two participants were first language (L1) Swedish speakers, two other participants were L1 Spanish speakers, one was a Dutch speaker, and another was a German speaker. Group 1 and group 2 consisted of students in Cognitive Science, while group 3 consisted of two academics in Language Teaching and Learning, and Educational Work. No information about age or gender was collected during this study.

Instruments and Materials

GenoGraphiX-Log

The participants first wrote a short academically styled text about the potential impact of Artificial Intelligence on the future of teaching, writing, and grading in the keystroke logging program GenoGraphiX-Log2 (GGXLog2). While writing in this program, each keystroke is registered, allowing for the analysis of markers of increased or decreased cognitive effort, including but not limited to cognitive pauses (hereafter referred to simply as pauses and defined as any pause equal or greater than two seconds), deletions (any deletion of text), bursts (sequences of keystrokes between interruptions such as pauses or deletions). The participants either downloaded the program themselves or wrote on a computer provided by the researcher. The writing began by starting a new free-writing session and pressing record, which unlocks the writing area, making it possible for the participant to start writing.

InstaText

InstaText is an AWE system powered by AI that suggests improvements to the text input and was deemed suitable for the purposes of this study because of its focus on academic and L2 writing, including its usability, as previously outlined. The original text is either pasted

or written in a window on the left side of the screen. Suggested improvements are presented to the right after clicking on a button. The added text is highlighted with a green underscore within the original text, and suggested removals are red and crossed out. These suggestions can be clicked and then accepted by pressing the check mark or rejected by pressing the cross. Users can also write new text in the right-hand window among the suggestions, and therefore substitute the AWE suggestions with their own. However, to get new suggestions on the updated text the user must go through the whole text and regenerate suggestions. The suggestions made by InstaText are not visually differentiated depending on its relation to spelling, grammar, clarity, and so on—that is, every suggestion is displayed in the same way, and it is up to the user to interpret what change the suggestion would cause. Suggestions can be as small as the placing of a comma or consist of an entirely rewritten sentence.

Text Inspector

To compare the differences between the original and revised text, an AWE called Text Inspector was used (Text Inspector, n.d.-a). Text Inspector analyzes text for purposes such as preparing teaching materials, tests, and evaluating student's work and provides measures of language quality, lexical diversity, sophistication of vocabulary, readability, and more. For the purposes of this study measures of textual lexical diversity (MTLD), Flesch Reading Ease, and Text Inspector's own language quality metric were selected to measure complexity, text readability, and language quality before and after revision with InstaText.

Procedure

The task was split into two stages, one initial writing stage and one revisions stage, the latter being promptly followed by a stimulated recall interview. The writing stage was carried out either at home on the participants own computers, or on the researchers' computers (if the participants were unable to download and use GGXLog2 on their own computers). Participants were given detailed instructions on how to download and use GGXLog2. A date and time for writing was decided, where the writing description and prompt were sent 10 minutes before to limit the preparation times for the participants. Participants were instructed to write for approximately one hour and were allowed to fall short or exceed the time frame by a reasonable amount if they could not write more or if they wanted to finalize their text. They were further instructed to write the text in English and to employ the same tone they would use in an academic paper. They could use other sources, such as the internet or articles, as they wished and did not have to properly cite. Furthermore, no spelling or grammar checkers were allowed. When they were done the keystroke logging file was collected.

The next stage of revising their short text was done as soon as possible after the writing stage, and the data for this stage was collected by the researcher. Their text was pasted into InstaText to generate suggestions. The participants could then start revising their text by accepting, rejecting, or substituting the suggestions made by the AWE. The participant's screens were recorded while interacting with the AWE. Immediately after they were finished with their revisions, the participants were asked to watch their interactions and discuss their thought process behind their decisions and the AWE suggestions in a semi-structured stimulated recall interview, while being recorded by the author.

Analysis

Keystroke Logging Analysis

Keystroke logging data was extracted from GGXLog2, including session length (amount of time spent writing), text length, pause percentage of session, mean pause times,

pause and deletion frequency. Descriptive statistics (mean, standard deviation) were calculated for these variables to provide an overview of the writing process for each participant and group to facilitate comparison between the written text and the participants' data from their interaction with InstaText. The relationship between keystroke metrics (e.g. pause frequency, deletion frequency), text characteristics (e.g. text length, lexical diversity), and interaction data (e.g., suggestion frequency) were explored through simple data visualizations to understand how cognitive effort during writing associated with the complexity and quality of their texts, as well as the connection between markers of demand (such as pause frequency) and the AWE suggestion frequency—for the purpose of investigating RQ1.

Text Analysis

Text Inspector (Text Inspector, n.d.-a) was used to provide measures of text complexity, readability, and quality, as these measures are related to writing ability and therefore, the participants' proficiency. The Measure of Textual Lexical Diversity (MTLD) was used to determine the variety of vocabulary exhibited in the different versions of the text. It calculates the mean length of word strings in the measured text while maintaining a given Type-Token Ratio (TTR)—that is, the unique words divided by the total number of words. The measured text is evaluated sequentially, and when the TTR of a string of words falls to 0.72 the process resets and starts to evaluate the next string. This measure of lexical diversity is used to ensure that the result is not affected by the length of the measured text, as is the case with other measures (McCarthy & Jarvis, 2010). Lexical diversity can indicate text complexity, as different mean segment lengths can be associated with various stages of language development (Malvern et al., 2004). Another measure that targets complexity and readability that was used in the current thesis is Flesch Reading Ease (Flesch, 1979). In this measure, the ratio of total words, phrases, and syllables are used to calculate a score between 0-100—where 0-50 indicates complex texts (college level and above). The overall language quality was measured using Text Inspectors own measure of language quality, calculated from a range of metrics, including number of syllables, lexical diversity, and lexical sophistication, among others (Text Inspector, n.d.-a). This yields a score from 0-100% (100% indicating the level of a native speaker), including an estimation of the language level according to the Common European Framework of Reference (CEFR).

These measures were used to conduct a comparative analysis through data visualizations. The text analysis was made on the participants' revised texts (the output from InstaText) as well, not only the original texts, to get a deeper understanding of their interactions with the AI powered AWE system—as their interactions with InstaText would impact these measures. For example, if participants note that they wanted to accept certain types of changes, those changes might be reflected in the analysis of the revised text. The text analysis of the revised texts can then be compared with the stimulated recall analysis and interpreted through Hayes and Berninger's model. This relates with the goal of RQ2 and its investigation of the participants' own understanding of their decision-making, and the investigation their behavior in terms of Hayes and Berninger's (2014) cognitive model of writing as posed by RQ3.

Interaction Analysis

The number of suggestions, suggestion frequency, and decisions (accept, reject, substitute), including type of decision and decision time, were manually extracted through the recordings of the interactions with InstaText. Acceptance rates were then calculated for each participant and group. Descriptive statistics of the decision times for each participant were analyzed to understand the cognitive load associated with different types of decisions and the interaction data was also compared with the keystroke logging data to investigate RQ1.

Stimulated Recall Analysis

Because RQ2 inquired into the participants' own understanding of their decisions, and the stimulated recall interviews consisted of a large amount of data from semi-structured interviews, thematic analysis was deemed suitable for the purpose of the study. The analysis was conducted according to the methods of Braun and Clarke (2021), with necessary adaptations based on the nature of the stimulated recall interviews. Due to time constraints, no additional and independent coders were used; therefore, the current study undertook a reflexive approach to thematic analysis as outlined by Braun and Clarke. With the absence of additional coders, inter-rater reliability could not be measured. However, the lack of reliability testing and multiple coders is in line with the qualitative paradigm of reflexive thematic analysis, emphasizing a subjective, interpretive reflective process and iterative theme development.

After data collection, the interviews were transcribed and carefully examined to identify comments exhibiting clear metacognitive reflections by the participants. Comments that exhibited such reflections were extracted and coded according to their content. The codes were then used to identify repeating patterns across participants, and excerpts that received similar codes for multiple participants were categorized according to the codes into separate potential themes representing types of metacognitive reflections about AWCF decisions. These initial themes were then reviewed, similar themes were merged, and themes that were deemed to not accurately reflect the data were discarded. Each reflection type was noted for its presence or absence across participants. Lastly, the qualitative insights from the stimulated recall analysis were compared with the quantitative metrics to provide a comprehensive understanding of participants' behavior and thought processes.

Ethical Considerations

Participants were informed about the purpose of the study and that participation was voluntary and could be terminated at any time, which would result in the deletion of their data. Each participant gave their consent for their data to be used. In addition, their data was anonymized so that it could not be traced back to them and no personal information was collected during the writing and use of the AWE system. To preserve anonymity, the participant's respective L1 is not presented, since it could be used by some, together with other information, to identify participants. The stimulated recall interviews were recorded and stored securely in Microsoft Teams, and once transcribed, the audio recordings were promptly deleted. All quotes from the interviews were edited to remove identifying comments.

Results and Discussion

The results and discussion section is split into three parts: first, the results of the quantitative analyses (keystroke dynamics, text analysis and interaction analysis) and a discussion of the quantitative results; second, the stimulated recall analysis; and third, an integrated discussion that interprets the cumulated findings through Hayes and Berninger's (2014) cognitive model of writing.

Quantitative Results and Discussion

Quantitative Results

The keystroke logging data, as well as data on the text and length of the writing session can be viewed in table 1. Participants spent on average 58 minutes and 40 seconds writing their texts, and texts were on average 4403 characters long. There was a significant amount of

variance in both measures. The bachelor students in group 1 had a difference in text lengths of only 593 characters, while their session times differ by about 21 minutes. Both participants in group 2 spent a longer time on writing than any other participants while also exhibiting divergent text lengths with a difference of 1763 characters. Participants 1 and 2 exhibit the most similar session times with a difference of seven minutes, while their text lengths differed with 1140 characters. Both the text lengths and session times displayed variation (session time with a SD of 16 minutes and text length with 1013 characters), but there is no indication of text lengths increasing as session time increases (i.e., longer writing time did not necessarily lead to longer texts).

Table 1

Original text data and keystroke logging data of participants.

Group	Participant	Session time	Text length (chars)	Mean Pause Time (s)	Pause Percentage of session	Pause frequency	Deletion frequency
G1	P3	00:35:13	3872	6.78	38.52%	3.10	15.19
	P6	00:56:23	3279	7.43	55.14%	7.65	20.13
G2	P4	01:10:35	3518	10.29	58.04%	6.79	29.16
	P5	01:21:45	5281	14.37	70.57%	4.56	18.05
G3	P1	00:57:44	5805	10.65	51.02%	2.86	14.88
	P2	00:50:24	4665	7.18	47.23%	4.27	12.50
Mean		00:58:41	4403.33	9.45	53%	4.87	18.32
SD		00:16:07	1013.28	2.92	11%	1.95	5.94

Note. "Pauses" refer to cognitive pauses, that is, pauses that are ≥ 2 seconds, and "Pause Percentage of session" refers to the portion of pauses in the session time. All frequencies are per 100 written characters.

The Text Inspector analysis of each participant's original and revised texts can be seen in table 2. Groups 1 and 2 show a decrease in text complexity for the original texts as measured by both lexical diversity and Flesch Reading Ease, with minimal variation between participants (see figure 1). Group 1 received a mean lexical diversity of 72.19 and group 2 received a mean lexical diversity of 83.71. For groups 1 and 2, the lexical diversity seems to rise with academic level. But this is not the case for group 3, which simultaneously has the participant with the lowest lexical diversity (participant 1) and the highest lexical diversity (participant 2), with scores of 62.12 and 101.6 respectively. As for Flesch Reading Ease, group 1 received a mean of 55.745 and group 2 received a mean of 44.4 with minimal variation between participants within the groups. However, even though the mean of group 3 decreases (43.24), the participants within the group show a lot of variation, including participant 2 with the lowest Flesch Reading Ease of 37.38, indicating the highest complexity, and participant 1 with a readability score between participant 6 in group 1 and participant 5 in group 2 (see figure 1). The data does not show increasing complexity together with academic level across every group because of the variation for group 3 in both text complexity measures.

The analysis of language level revealed a mean language quality of 72.91% for the whole sample, while group 1 received a mean of 70.08%, group 2 received 73.57%, and group 3 received 75.09%, with a small amount of variation (2.52%). This indicates that the overall language quality slightly increased with academic level. According to Text Inspector, every participant received a level of C2 (proficient), except for participant 6 in group 1 who received C1+ (advanced), indicating that each participant exhibited a high-level of English proficiency

(Council of Europe, n.d.). Therefore, the language proficiency was homogeneous across participants and groups.

Table 2

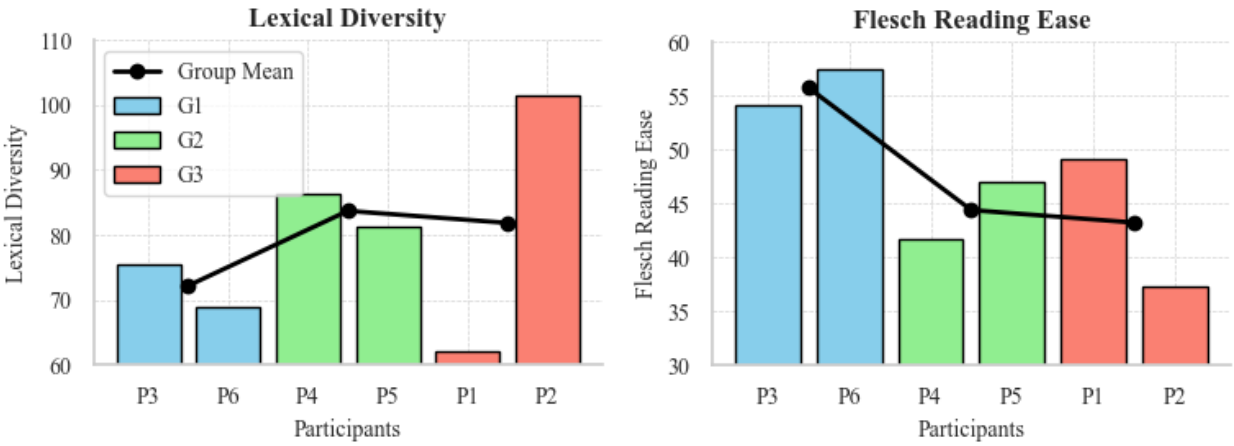
Text Inspector's analysis of the original text vs. the revised text.

Group	Parti-cipant	Typo frequency		Flesch Reading Ease		Lexical diversity (MTLD)		Text Inspector language quality	
		Original	Revised	Original	Revised	Original	Revised	Original	Revised
G1	P3	2.72	1.04	54.08	55.58	75.39	72.8	71.56%	71.72%
	P6	2.39	0.53	57.41	57.72	68.98	65.75	68.6%	66.67%
G2	P4	0.86	1.23	41.71	43.15	86.25	80.18	74.05%	75.47%
	P5	0.35	0.46	47.09	47.19	81.18	71.71	73.1%	72.84%
G3	P1	1.58	0.97	49.1	48.66	62.12	63.86	74.52%	73.33%
	P2	1.77	0.67	37.38	37.75	101.6	78.06	75.65%	74.4%
Mean		1.61	0.82	47.80	48.34	79.25	72.06	72.91%	72.41%
SD		0.90	0.31	7.48	7.50	13.90	6.48	2.52%	3.09%

Note. "Typo frequency" refers to spelling mistakes per 100 words, and "Original" and "Revised" refers to the original text and the revised text after using AWE, respectively.

Figure 1

Lexical diversity (left) and Flesch Reading Ease (right) of each participants' original text in each group.



The text length tended to decrease after the use of the AI tool InstaText, where the mean text length went from approximately 4403 characters in the original text to 4320 characters in the revised text. The lexical diversity also tended to decrease after the use of the AI powered AWE, going from a mean lexical diversity of 79.25 before AI, and 72.06 afterwards (table 2). Only participant 1 in group 3 marginally increased their lexical diversity from 62.12 to 63.86. Flesch Reading Ease saw very little change, exhibiting a minimal increase in score, going from a mean of approximately 47.8 to 48.3. The language quality of the revised texts exhibited a minor decrease, going from a mean of approximately 72.9% in the original texts to 72.4% in the revised texts, but this resulted in no change of the estimated CEFR level. Lastly, the typo frequency went down from a mean of approximately 1.61 for the original texts to 0.82 for the revised texts. The remaining typos could be a product of substitutions or additions to the text during the revision stage since InstaText did not evaluate the newly added text automatically.

This could also explain the increase in typo frequency for two participants—participant 4 went from a typo frequency of 0.86 in the original to 1.23 in the revised text, and participant 5 went from a typo frequency of 0.35 in the original to 0.46 in the revised text.

The data of the participants' interactions (i.e., the number of suggestions, accepts, rejections and substitutions) is presented in table 3, while the analysis of the decision times for the interaction along with acceptance rates can be viewed in table 4. The acceptance rate of the participants shows a variation of 11%, with participant 5 having the smallest acceptance rate (69%) and participant 2 having the highest acceptance rate (95%). Each group has one participant with a relatively high acceptance rate above the mean of 81% and one below. As such, this data can be formed into two separate groups that include one participant from each academic group each (see figure 2).

As can be seen in figure 3, there seems to be a relationship between acceptance rate, and lexical diversity—as lexical diversity increases, so does acceptance rate—with the exception of participant 5. This could indicate that the participants who wrote more complex texts in terms of vocabulary use were the ones that had an increased likelihood of accepting the suggestions generated by the AI. For acceptance rate and Flesch Reading Ease, the trend holds for each participant except for participant 1 and 5 who has relatively low acceptance rates in relation to their Flesch Reading Ease score (see figure 3).

Additionally, the decision times for accepting revisions were shorter than for rejecting suggestions, where the mean accept time was 5.29 seconds long and the mean reject time was 8.82 seconds long. Substitution times tended to be significantly longer, at a mean decision time of 20.8 seconds. It therefore took longer for participants to decide to reject a suggestion, than to accept it, and even longer to decide to substitute the AI's suggestion with their own revision. Accepting suggestions was also more consistently faster than rejecting or substituting.

Table 3

Total number of suggestions from InstaText and participants' decisions.

Group	Participant	Total suggestions	Accept	Reject	Substitute
G1	P3	141	118	23	0
	P6	114	84	26	4
G2	P4	116	104	9	3
	P5	149	103	40	6
G3	P1	201	142	57	2
	P2	157	149	8	3
Mean		146.33	116.67	27.17	3
SD		31.95	24.91	18.82	2

Table 4

InstaText interaction data for each participant.

Group	Parti-cipant	Suggestion frequency	Acceptance rate	Accept time (s)		Reject time (s)		Substitute time (s)	
				M	*SD*	*M*	*SD*	*M*	*SD*
G1	P3	3.65	84%	5.26	4.28	5.02	2.78	-	-
	P6	3.49	74%	5.48	4.3	10.29	11.66	8.65	4.17
G2	P4	3.32	90%	5.16	4.8	14.41	13.41	11.15	0.44
	P5	2.83	69%	4.17	3.24	3.99	2.97	10.24	3.06
G3	P1	3.54	71%	5.17	3.41	7.03	4.45	37.69	29.32
	P2	3.37	95%	6.5	5.84	12.16	12.01	36.28	32.33

Note. The hyphen ("-") indicates that the participant made no substitutions.

Figure 2

Acceptance rate for each participant and the mean acceptance rate.

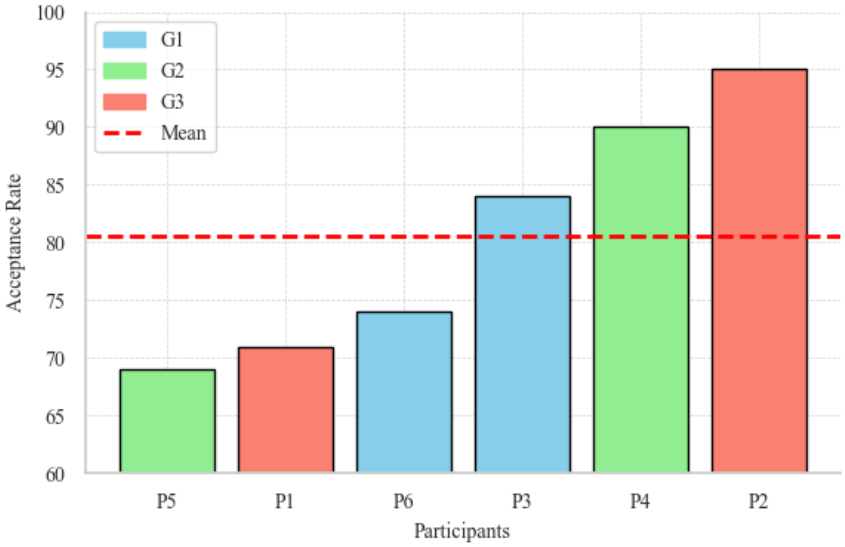
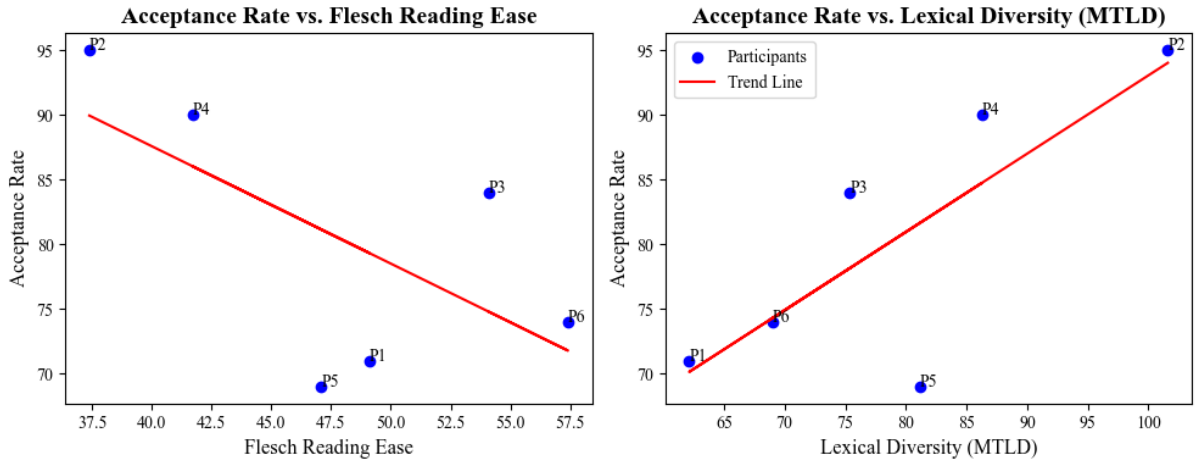


Figure 3

Relationships between acceptance rate, readability (left), and lexical diversity (right).



The interaction data also reveals a possible relationship between the frequency of suggestions (suggestions per 100 characters) and the readability of the texts, as texts with higher Flesch Reading Ease score tended to receive more suggestions from InstaText (see figure 4). There was a similar trend for lexical diversity, where participants who wrote texts with higher lexical diversity received less over all suggestions from the AI powered AWE tool. However, participant 5 distinctly deviates from this pattern with their significantly lower suggestion frequency, that could be due to their significantly lower typo frequency (0.35, see table 2) since the participants with lower typo frequency received fewer suggestions (see figure 5). In figure 5, a connection between mean pause length, pause percentage of the session, suggestion frequency and typo frequency can be observed in several different graphs—as participants spent more time pausing during the writing session, typos decreased, and those participants also received fewer suggestions from InstaText.

Figure 4

Relationships between suggestion frequency readability (left) and lexical diversity (right).

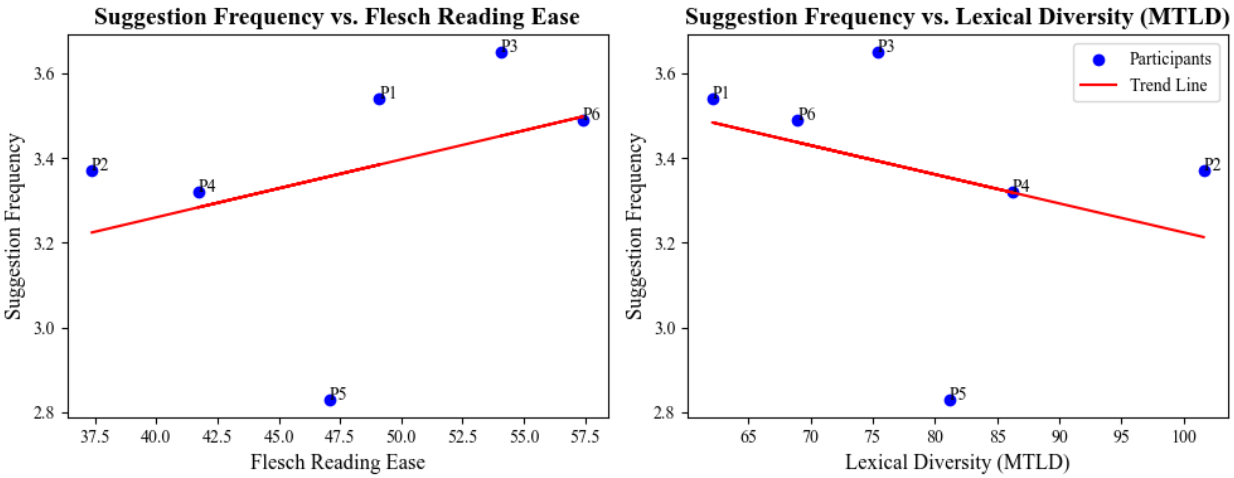
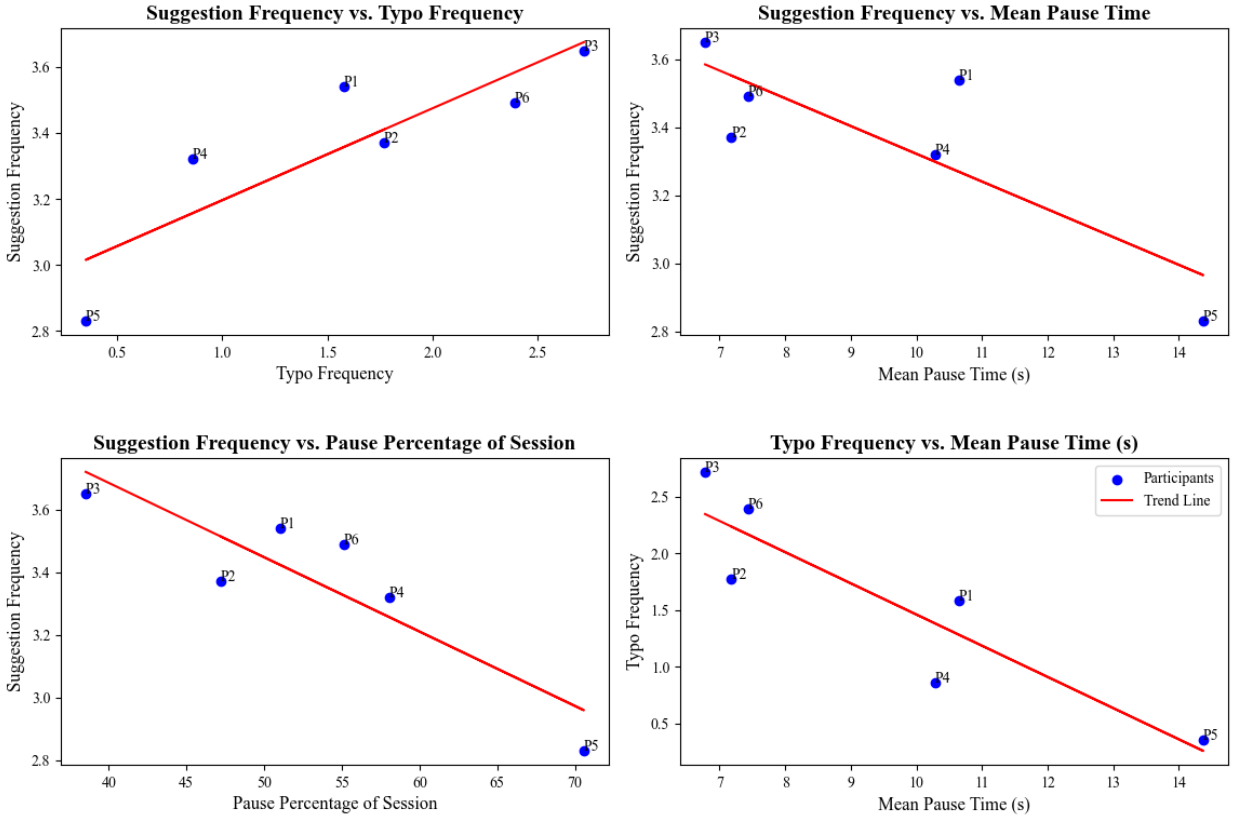


Figure 5

Relationships between suggestion frequency, typo frequency, mean pause time, and pause percentage.



Quantitative Discussion

The keystroke logging data provides an overview of participants' effort in writing their original texts. As could be seen, there was a noticeable variation in session times, text lengths and cognitive pauses among the participants. However, it was clear that longer session times did not necessarily lead to longer texts (see table 1). Some participants were more efficient than others, writing a lot in a short time-span—some might have spent a long time writing a short text, which could indicate a struggle with different processes such as planning their texts, or the retrieval of knowledge about the text's topic.

For the text analysis with Text Inspector, the measures of complexity and readability revealed some tendencies. Writers with higher academic experience seem to show more lexical diversity and complexity. However, the individual data was not consistent in the advanced academic group considering participant 1 in group 3 shows less than expected scores in both lexical diversity and Flesch Reading Ease. As previously discussed, Guo et al. (2022) found that students with lower writing proficiency tended to have a decreased acceptance rate compared to students with higher writing proficiency. The participants in the current study exhibited homogeneous language quality in their text in terms of CEFR estimate and Text Inspector's language quality measure (see table 2), while showing high variability in acceptance rates. Therefore, the data did not show any relationship between language level and acceptance. Furthermore, acceptance rates did not increase in relation to academic level (see figure 2). However, one important part of writing proficiency is the ability to utilize a diverse vocabulary as measured by lexical diversity, which rose together with acceptance rates (except for one participant; see figure 3). The comparison of the original and revised text revealed that

text became less complex in terms of vocabulary (see table 2). This could be the result of a tendency either of the participants to choose suggestions that use less diverse words or use more words consistently throughout the text. But it could also be a general preference of InstaText to replace certain terms consistently.

The data of the interaction with InstaText suggest that rejecting AWCF required more cognitive effort than accepting AWCF since mean decision times for rejecting suggestions were on average 3.53 seconds longer than for accepting suggestions. But the highest decision times were associated with substitutions. This is to be expected since substitutions require the production of text on the part of the participant instead of the generative AI—that is, instead of only evaluating and accepting or rejecting already composed text, the participant must first evaluate the AWCF, then go through the cognitive processing associated with text production. This would therefore require the engagement of the proposer, the translator, the transcriber, and the evaluator, instead of only the evaluator according to Hayes and Berninger's (2014) cognitive model of writing.

A significant observation relating to RQ1 is the relationship between suggestion frequency and mean pause time as visualized in figure 5. It seems that there was an identifiable relationship between the keystroke metrics and AWE suggestions to some degree. Participants with higher mean pause times received less AWCF, suggesting that certain cognitive engagement produced noticeable artifacts in their text, which the AI detected. On closer inspection of the data, it seems that the typo frequency could be the common denominator, as the analysis reveals a relationship between pause time (both mean pause time and pause percentage of the session) and suggestion frequency—as typos increase, so does the number of suggestions, and as pause time increases, the number of typos decrease. It is reasonable to assume that what the relationship between increased pause time and a decrease in typos indicates is a connection with the evaluation of texts. In this case, the evaluator in the process level of Hayes and Berninger's (2014) cognitive model evaluating the spelling of the transcriber's output. Participants with lower pause times probably spent less time evaluating the spelling of their texts, leading to more typos and a higher suggestion rate.

It would not be unreasonable to expect a relationship between suggestion frequency and deletion or pause frequency, as these metrics have been associated with increases in cognitive demand—as the demand of the task exceeds working memory capacity, pauses and mistakes increases (Conijn et al., 2019; Wengelin, 2006). One could then expect InstaText to notice such mistakes, which would lead to more suggestions. Nevertheless, no keystroke logging metrics other than mean pause length and time spent pausing was related to suggestion frequency, and because of the relation to typo frequency, this relationship mostly seems to be associated with surface level evaluations of grammar and spelling. To see if there would be any connection between the measures that keystroke logging offers and AWE suggestions for deep level evaluations such as meaning construction, controlling typos using automated spelling correction when writing the original text would be pertinent. A higher resolution in the keystroke logging analysis can also facilitate this goal by providing insight into where the pauses were made—for example, word level, sentence level, paragraph level. This would shine further light on what processes the pauses were associated with, since pauses at different text levels have been associated with different types of processing in L2 writers: pauses at sentence level tend to be related to higher level planning processes, and word level pauses tend to relate to the translation of ideas into language (Galbraith & Vedder, 2019).

Stimulated Recall Analysis

Six main metacognitive reflections were identified in the participants' reasoning behind their decisions. Table 5 demonstrates each reflection type and which participants demonstrated

what reflection. The analysis revealed six main reflections in the stimulated recall sessions. Each reflection was demonstrated by every group in at least one participant. These reflection types are by no means exhaustive, they merely reflect what participants chose to discuss in detail, as the analysis focused on the excerpts that demonstrated deeper metacognitive engagement.

Table 5

Each type of metacognitive reflection demonstrated by each group and participant.

Reflection	G1		G2		G3	
	P3	P6	P4	P5	P1	P2
Sense-preservation	X	X	X	X	X	X
Conciseness and complexity		X	X	X	X	X
Language influence	X	X	X		X	
Formal language use	X		X		X	X
Word consistency	X	X	X	X		X
Automatic acceptance	X	X	X		X	X

The first and by far most prevalent reason, demonstrated by each participant (see table 5), was accepting or rejecting AWCF based on *sense-preservation*, that is, preserving the intended meaning of the texts. The most common decision associated with this reflection was the rejection of suggestions that were perceived as changing the intended meaning of their text, as demonstrated by participant 3 in the following example:

Stimulated Recall		
AWCF	**Original text**	**Comment**
fine-tuning —> tweaking	With the correct script, the correct fine-tuning and the correct prompt, a LLm could grade an essay under 1 minute	Here it completely misses the meaning if I accept the change. Yes, because "fine-tuning" is something different than "tweaking". It might not have the whole context of the document.

Here, participant 3 rejected the suggestion to change “fine-tuning” to “tweaking” because of discrepancy between the alteration and the participant’s intended meaning.

Sense-preservation also included accepting suggestions that fell more in line with what participants failed but wanted to express—where the AWCF nor their own texts captured the intended meaning—leading to substitutions. For example, when participant 6 decided to rewrite their sentence after seeing InstaText’s suggestion and their own text:

Stimulated Recall		
AWCF	**Original text**	**Comment**
which to much —> as too many do —> be made	When doing this, it is important not to loose your opinion and tone of the material, which to much changes can do	Right, that was a weird sentence. It’s probably weird since it was a bit weird before, but I think it wanted to make changes that made it miss the meaning. Yeah, then I think I rewrote it a bit.

Here, participant 6 identifies a discrepancy in meaning between their own text, “which too many changes can do”, and the InstaText revision, “as too many changes can be made”.

It would be reasonable to assume that the participants' language level, to some degree, determines their success rate in identifying meaning changes in the suggestions, especially the more subtle varieties. To what degree the participants successfully identified sense-alterations in texts cannot be identified here due to the nature of the analysis and the data—the variety of language ability is far too narrow, as seen in the Text Inspector language quality measure and CEFR estimation. Future research on people's propensity to accept meaning changing alterations would shine further light on the risks of AWE usage by writers of varying language skills.

The second type of metacognitive reasoning regarded the *conciseness and complexity* of either their original texts or the AWCF. This reflection was demonstrated by all participants except one, and included text alterations based on unnecessarily complex sentence structure and a perceived conciseness of the AWCF or the participants' own texts, as demonstrated in the following example:

Stimulated Recall		
AWCF	**Original text**	**Comment**
in Higher Education, stressing the impossibility of excluding —> into higher education and emphasise that it is impossible to exclude	Others are more likely to see the potential benefits of integrating AI in Higher Education, stressing the impossibility of excluding technology from modern teaching and learning.	It felt like it suggested more effective solutions, that is, it becomes more dense instead of my complicated wording.

Here participant 2 accepted the AWCF, and expressed the opinion that InstaText was more effective in its word usage and made their writing less complicated. The quantitative data suggests texts became marginally more readable, therefore less complex, as demonstrated by the slight increase in mean Flesch Reading Ease for the revised texts (47.80 to 48.34; see table 2). However, since this was such a small increase, it might not have an identifiable impact on readability.

The third metacognitive reasoning yielded by the analysis of the stimulated recall sessions was *language influence*—reflections on the influence of languages in their language use. This was interrelated to other reflections such as the previous conciseness and complexity, as demonstrated by both Spanish speaking participants, where they noted the tendency to use more words than necessary and an increased use of commas. The following example demonstrates this reflection:

Stimulated Recall		
AWCF	**Original text**	**Comment**
to be —> write productive in their writing areas —> productively and assist them in areas aspects —> which	AI tools can help students to be more productive in their writing and be an aid in aspects they are less proficient.	That whole suggestion sounded a lot more concise and precise. Sometimes because my first language is Spanish, I tend to say things with a lot of words and, yeah, in that case it helps to be more concise.

The participant articulated their tendency to use an unnecessary number of words when writing, which they attributed to the influence of their L1. This reflection was mirrored by another participant, where they also note that their supervisor has commented on their tendency to use

a lot of commas in their writing, also attributed to L1 influence. This reflection was observed in four of the six participants since every participant wrote in L2 English. Both Swedish speaking participants also noted L1 influence, in the form of the accidental usage of "Swenglish" vocabulary(i.e., the ungrammatical influence of Swedish on English vocabulary).

One of the task requirements was to write the text in formal English, in academic parlance. This led to reasoning about *formal language use*. Thus, four of the six participants explicitly reflected on their texts' and the AWCF's formality. In one participant, views on the sophistication of English variants influenced their opinion about the suggestion's formality, where participant 3 rejected a suggestion based on it sounding more like US English, which they perceived as less academic than UK English:

Stimulated Recall		
AWCF	**Original text**	**Comment**
many issues —> a lot of problems	But this created <u>many issues</u>.	I try to write in British English and according to me "issues" sounds more British, and it also feels more proper when it is supposed to be academic.

This resembles the language influence reflection, however, here the language influence is from a variant of the target L2. Participant 4 also noted the preference of using UK English in their writing based on it being the norm in their institution.

In the quantitative data, there is a clear drop in the lexical diversity in the revised text (from a mean of 79.25 to 72.06; see table 2) indicating that the vocabulary of the revised texts became decreasingly varied. This is reflected in explicit reasoning by the participants. In five out of six participants, *word consistency* was mentioned in their metacognitive reflections, indicating a preference for using the same word for the same reference. Participant 5 demonstrated this reflection after rejecting the AWCF since they preferred using "implications" because of its previous use in the text:

Stimulated Recall		
AWCF	**Original text**	**Comment**
implications —> consequences	Aside from the negative <u>implications</u>, the use of AI in education can have various positive impacts.	I think I rejected that it should be "consequence" and I kept it "implications" because I did that in the rest of the paper as well.

Multiple participants indicated a preference for suggestions where the AI consistently suggested exchanging certain words—for example, suggesting "as" instead of "since" and "such as" instead of "for example"—as the lexical diversity of the texts also indicates. However, suggestion frequency decreased as lexical diversity increased (see figure 4). This suggests it might have to do with participants own preferences rather than a propensity for the AI to suggest consistent word use, as does participant 5's previous reflection. However, several participants (demonstrated here by participant 2) noted consistent AWCF for the same phrases:

Stimulated Recall		
AWCF	**Original text**	**Comment**
e.g. examination —> assessment	Many see the potential risks in the features offered by AI and suggest a return to previous forms of e.g. examination	It suggested "assessment" through the whole text instead of "examination" and then I thought yeah if that is a word that fits and is used more frequently then I'll take it.

Participant 2 noted that InstaText consistently exchanged “examination” with “assessment” throughout the text.

Looking at the interaction data (table 4) accept times are, on average, -3.53 seconds shorter than rejected times. It therefore seems to take more cognitive effort to decide to reject, than to accept, AWCF suggestions. This is also displayed in participants metacognition since a common reason for accepting suggestions was based on simple heuristics, leading to the *automatic acceptance* of AWCF, exemplified in the following stimulated recall:

Stimulated Recall		
AWCF	**Original text**	**Comment**
n —> n, a —> an	Education is ripe for innovation and is based on a outdated model from the early industrialisation of society.	Some of them where just feeling, but everything relating to grammar made me accept.

Participant 3 states that if they identify the origin of the suggestion as related to spelling or grammar, such as the addition or removal of a comma or the substitution of “a” to an “an”, they automatically accept the suggestion. It seems therefore that there is a level of trust offered to the AI’s grammatical capabilities that is not offered to that of sense-preservation. Participant 6 confidently states in a later comment that they “know a lot about the structure of such systems and my own capabilities”, and therefore offers full trust to the system when it comes to grammatical changes. Suggestions related to spelling and grammatical errors would have significantly impacted the average decision times of participants’ interactions with InstaText since the original text was written in GGXLog2 without any AWE features. To investigate cognitive effort in relation to AI powered AWE and its effect on decision times, eliminating grammatical and spelling errors would be pertinent in future research.

Participants were also unsure of some of the AWCF, especially when they could not determine any meaningful difference between their own text and the AWE’s suggestion. This could sometimes lead to automatically accepting the AWE suggestion, as was the case with participant 6:

Stimulated Recall		
AWCF	**Original text**	**Comment**
producing their —> creating	AI should not be used for producing their teaching materials	There are some when I feel like both work as well but then it might often be the case that I accept the suggestion. I might think that the suggestion works as well or that it has better grammatical knowledge and then I will accept.

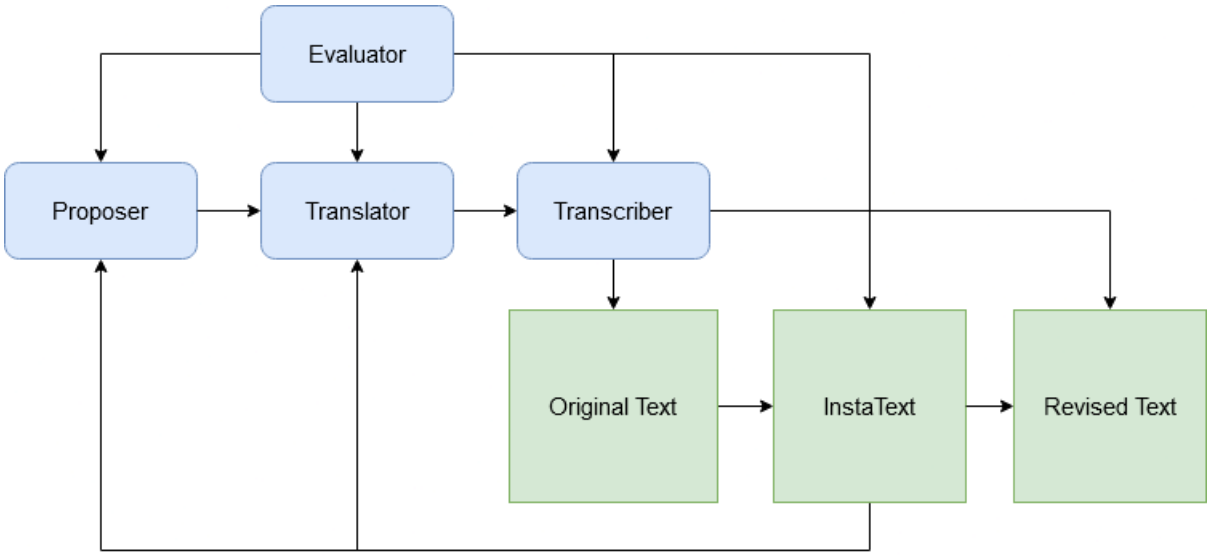
When faced with doubt, the participants did not always accept the decisions, but made their choice based on aesthetic preferences.

Integration with a Cognitive Model of Writing

The discussed reflections from the stimulated recall are metacognitive accounts of the writing process when using AI powered AWE. As such, these accounts can be interpreted in terms of the writing process as outlined in Hayes and Berninger's (2014) cognitive model of writing along with the behavioral data. AI powered AWE was used as a tool for revision, and Hayes and Berninger (2014) view revision not as a writing process but as a specific writing task, a task in which the writing processes are used to replace the originally written text. Writing with AI-assistance for the revision of text can therefore also be considered as a specific writing task that consists of relations between writing processes and the AWE technology. AWE systems such as InstaText can be viewed as a digital form of the collaborator or critic within the task environment of Hayes and Berninger's model. But it would be reasonable to assume that the way in which interactions between a writer and a human collaborator would significantly differ from the interaction between a writer and an AI powered AWE. To clarify the specific relations between the writing processes exhibited in the current study and InstaText, a diagram of the relationships within the process level can be viewed in figure 6.

Figure 6

Diagram of the relationship between cognitive writing processes according to Hayes and Berninger's (2014) model and the AI AWE InstaText.



Note. Blue rectangles with rounded edges represent writing processes and green squares represent the task environment. The arrows represent relations between components. Not every component or relation is depicted and everything that is depicted in the diagram falls within the process level.

Hayes and Berninger (2014) states that the evaluator can interact with any of the other cognitive processes (the proposer, translator, and transcriber) as well as the text written so far; when writing with AI powered AWE, the evaluator also interacts with the AWE output. The main relationship is between the evaluator and the AWE (InstaText). In this relation, the evaluator evaluates the output of the AWE and decides whether to accept, reject or substitute. The AWE takes the original text as input and generates suggestions. If the evaluator accepts the decision, the output replaces the input in the revised text, if it rejects it, the input remains (both are represented as the relation between the AWE and the revised text). If the evaluator

deems both the original text and suggestions as insufficient, a substitution is made and the writing process either starts at the proposer if a new idea is put forth, or the translator reconverts the idea into a new linguistic string.

Reflections on sense-preservation could be interpreted as metacognitive accounts of a specific evaluation process. This process involves comparing the idea generated by the proposer (the intended meaning) with both the meaning of the original text and the AI-generated suggestions. If the idea aligns with the originally produced text but not the suggestion, the AWE feedback is rejected; if it aligns with the suggestion but not the original text, the feedback is accepted; if neither, substitutions occur, and the writing process restarts.

The other metacognitive reflections are found within the bounds of the control level, that is, what controls operations of the process level (Hayes & Berninger, 2014). Participants' reflections on simplifying sentence structures or maintaining complexity reflect the work of the translator (converting ideas into linguistic forms) and the transcriber (converting these forms into written text). With AWE technology, this becomes an evaluation process of both the transcriber's output and the AI-generated suggestions. What determines the decision (accept, reject or substitute) is the alignment with participants preferences—in the case of conciseness and complexity, in the preference for straightforward language, as exhibited in the low lexical diversity of the revised text and participant's comments during stimulated recall. The preferences are found in the goals of the planner or genre knowledge of the writing schemas. The planner might set subgoals in the form of the structure of the text, while the genre knowledge includes the desired properties of the text. Thus, reflections on formal language use, word consistency, and language influence indicate the application of writing schemas or planning at the control level. The acceptance of suggestions that improve formality, word consistency, or align with specific language norms could reflect the strategic deployment of genre knowledge.

Finally, the automatic acceptance of certain types of suggestions could be the result of a delegation of the evaluation of surface level errors (such as grammatical mistakes) to the AWE to conserve cognitive resources for the evaluation of deep-level errors such as meaning, coherence, or clarity. Therefore, the reflection of this type of behavior also relates to the control level of Hayes and Berninger's (2014) cognitive model, as the control level is where regulatory factors direct cognitive processes to optimize resource allocation. The control level determines the evaluation through heuristics—which fit within the strategic knowledge of the writing schemas—and therefore the decision-making regarding the AWCF.

The relationships outlined above are by no means exhaustive, and different interpretations of the AI-assisted revision process are probably possible. Future research will be needed to sufficiently outline how cognitive processes operate together with AI-assisted writing process.

Concluding Remarks

For the purposes and scope of this exploratory study, the research questions have been answered. The first research question asked if there was an identifiable relationship between keystroke dynamics and the suggestions generated by AWE powered by generative AI. Participants who paused longer received fewer suggestions by the AWE system. This was likely due to the engagement in evaluation processes, resulting in fewer surface level mistakes such as typos, which caused fewer suggestions. The second research question asked how second-language academic writers understand AWE suggestions, and what their reasons were for accepting, rejecting or substituting those suggestions. Participants reflected on various aspects of their decision-making process. They noted trying to preserve their meaning throughout the revision process and convey their point as effectively as possible. They noted

influences of their second languages and evaluated the suggestions according to their formality and consistent use. Participants also explained that they sometimes automatically accepted suggestions if they saw them as relating to grammatical corrections. The third research question asked how AI powered AWE systems interact with the cognitive writing process as outlined by Hayes and Berninger's (2014) model of writing. The AI-assisted revision process was interpreted as a writing task that primarily consisted of a relation between the evaluator process and the output of InstaText. Different goals and preferences can affect the evaluation, as well as the suggestions alignment with the intended meaning as put forward by the proposer. The AWCF can also cause the writing process to start over, either at the proposer when new ideas might be put forth, or at the translator to try to reconstruct the intended meaning.

There are limitations in the current study. For one, the English proficiency of the participants was not formally measured. Future studies on AI-assisted L2 writing should measure English proficiency using standardized testing for a more comprehensive and reliable measure. Additionally, this was an exploratory study, investigating AI-assisted writing in a small sample of participants. Therefore, the results are preliminary by nature and interpretations of the quantitative and qualitative data should be treated with caution since multiple interpretations of the results are likely possible. This is especially the case of the integration of AWE technology within the Hayes and Berninger's (2014) cognitive model of writing since the specific role of AI powered AWE and its effect on cognitive processes was not experimentally tested. No advanced statistical methods were used to explore the relationships between measures. This was due to the low sample size which would limit the statistical power and potentially lead to misleading results. Therefore, the focus was on the qualitative insights and its comparison with descriptive statistics via visualizations to offer preliminary insights of the cognitive processes involved in AI-assisted writing. To build on the tentative results reported in the current study, future research should strive to utilize more representative samples and rigorous statistical approaches. Furthermore, no independent coders were used for the qualitative analysis, and no inter-rater reliability could therefore be measured and consistency in theme development could not be ensured. This can be addressed in future research through other approaches to thematic analysis that use codebooks and multiple coders (Braun & Clarke, 2021).

Another limitation and something future research should improve upon lies within the methodology. In the first stage of the task, participants wrote a text using GGXLog2 without using any form of AWE and no automated evaluation of spelling was offered to participants because of this. This led to a significant increase in the total number of suggestions. The analysis of the interaction data became more complicated as a result. For example, enumerating each reason for every decision when the mean number of suggestions was 146 across the six participants would have taken an infeasible amount of time for the scope of the current study. Future research should therefore integrate automated grammar checkers when writing the original text. This, together with a detailed keystroke logging analysis of the text level of pauses would allow the exploration of the connection between markers of cognitive effort and the construction of meaning in text rather than surface level evaluations of spelling. Additionally, since the participants displayed reflections on the impact of their first languages on their decisions, future studies could further assess the impact of first languages on decision-making during the use of AWE systems by including a larger sample of participants with different first languages.

References

Beikian, A., & Ganji, M. (2022). Efficacy of Instatext for improving Persian-English freelance translators' language quality: From perception to practice. *Journal of Foreign Language Teaching and Translation Studies, 7*(4), 59-86. https://doi.org/10.22034/efl.2022.368332.1205

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology, 18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Caporossi, G., Leblay, C., & Usoof, H. (2023). GenoGraphiX-LOG (Version 2.1.0) [Computer software]. HEC Montréal & University of Turku. https://ggxlog.net

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

Council of Europe. (n.d.). Global scale - Table 1 (CEFR 3.3): Common reference levels. Common European Framework of Reference for Languages. Retrieved May 19, 2024, from https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12402-3

Flesch, R. (1979). *How to write plain English: A book for lawyers and consumers*. Harper & Row.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering, 66*, 111–126. https://doi.org/10.1007/s12599-023-00834-7

Galbraith, D., & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In E. Lindgren, & K. P. H. Sullivan (Eds.), *Observing Writing: Insights from Keystroke Logging and Handwriting* (pp. 306–325). Brill. https://doi.org/10.1163/9789004392526_015

Galbraith, D., & Vedder, I. (2019). Methodological advances in investigating L2 writing processes: Challenges and perspectives. *Studies in Second Language Acquisition, 41*(3), 633–645. https://doi.org/10.1017/S0272263119000366

Grammarly. (n.d.). AI writing assistance. Grammarly. Retrieved June 11, 2024, from https://www.grammarly.com/ai

Guo, Q., Feng, R., & Hua, Y. (2022). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing? *Computer Assisted Language Learning, 35*(9), 2312–2331. https://doi.org/10.1080/09588221.2021.1879161

Hayes, J. R., & Berninger, V. W. (2014). Cognitive processes in writing: A framework. In B. Arfe, J. Dockrell, & V. W. Berninger (Eds.), *Writing development in children with hearing loss, dyslexia, or oral language problems: Implications for assessment and instruction*. Oxford Academic. https://doi.org/10.1093/acprof:oso/9780199827282.003.0001

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports, 13*, 18617. https://doi.org/10.1038/s41598-023-45644-9

InstaText. (n.d.-a). Academic writing. InstaText. Retrieved June 13, 2024, from https://instatext.io/academic-writing/

InstaText. (n.d.-b). How InstaText differs from grammar checkers. Retrieved April 20, 2024, from https://instatext.io/how-instatext-differs-from-grammar-checkers

InstaText. (n.d.-c). InstaText: Interactive text improvement tool [Software]. https://instatext.io/

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, 100450. https://doi.org/10.1016/j.asw.2020.100450

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

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*, 381–392. https://doi.org/10.3758/BRM.42.2.381

Mhlanga, D. (2023). Open AI in education: The responsible and ethical use of ChatGPT towards lifelong learning. In *FinTech and artificial intelligence for sustainable development: The role of smart technologies in achieving development goals* (pp. 387-409). Cham: Springer Nature Switzerland.

Shaw, C., Yuan, L., Brennan, D., Martin, S., Janson, N., Fox, K., & Bryant, G. (2023, October 23). AI in higher education: Fall 2023 update time for class study. Tyton Partners. Retrieved from https://tytonpartners.com/app/uploads/2023/10/GenAI-IN-HIGHER-EDUCATION-FALL-2023-UPDATE-TIME-FOR-CLASS-STUDY.pdf

Text Inspector. (n.d.-a). Text Inspector. Text Inspector. Retrieved May 18, 2024, from https://textinspector.com/

Text Inspector. (n.d.-b). TU Lexical Profile. Text Inspector. Retrieved May 18, 2024, from https://textinspector.com/help/tu-lexical-profile/

Wengelin, Å. (2006). Examining Pauses in Writing: Theory, Methods and Empirical Data. In K. P. H. Sullivan, & E. Lindgren (Eds.), *Computer Keystroke Logging and Writing: Methods and Applications* (pp. 107–130). Brill. https://doi.org/10.1163/9780080460932_008