Castledown

OPEN ACCESS

Technology in Language Teaching & Learning
ISSN 2652-1687
https://www.castledown.com/journals/tltl/

Technology in Language Teaching & Learning, 6(1), 1–19 (2024)
https://doi.org/10.29140/tltl.v6n1.1168

Adopting ChatGPT as a Writing Buddy in the Advanced L2 Writing Class

CAROLA STROBL[a]

IRYNA MENKE-BAZHUTKINA[b]

NIKLAS ABEL[c]

MARIJE MICHEL[d]

[a]*Department of Applied Linguistics, Translation and Interpreting Science, University of Antwerp*
carola.strobl@uantwerpen.be

[b]*European Languages and Cultures, Groningen University*
i.a.menke-bazhutkina@rug.nl

[c]*European Languages and Cultures, Groningen University*
n.abel@rug.nl

[d]*European Languages and Cultures, Groningen University*
m.c.michel@rug.nl

Abstract

Since its release, ChatGPT has raised concerns in many teaching contexts given its threat to reliably evaluating learners' knowledge and skills. Within task-based pedagogy, however, this technology opens new avenues for second language (L2) teaching when adopting the technology as a writing buddy. Our study explores how ChatGPT as a model impacts the revision process of advanced L2 writers of German. Twenty-two university students participated in a two-week classroom-based intervention, producing two summaries of popular-scientific texts in L2 German. After writing a first draft, they compared their summaries with texts produced by ChatGPT (3.5) and revised, where necessary, their own text. In this paper, we analyze all students' rubrics-based ratings of the ChatGPT models and present data of six focus students' screen-recorded revision processes that we coded for revision focus, source, and success. Results reveal students' growing awareness of characteristics of ChatGPT-output, such as linguistic accuracy and fluency, as well as its flaws in content provision. Revision data demonstrate that students skillfully made use of the models to improve their own texts. Our study provides evidence that using ChatGPT as models in writing and revision processes can stimulate higher-order thinking in the revision process of advanced L2 students.

Keywords: ChatGPT, model-based revision, L2 synthesis writing, inner feedback

Introduction

Since its release, ChatGPT (Open-AI, 2022) has raised concerns in education. Students at any level are tempted to use the tool extensively, while experts warn that the technology at times produces (easy-to-detect) nonsense. Not just in higher education, it is therefore crucial to foster students' awareness of its benefits and pitfalls. Inspired by translation pedagogy, we argue that adopting ChatGPT as a writing buddy during text revision processes may fulfil this awareness-raising function and foster the development of second language (L2) writing skills. Machine translation (MT) software based on large language models (LLMs) has been around for more than a decade (e.g., DeepL). It produces increasingly sophisticated output, often outperforming students translating into their L2. Since these tools are also widespread in industry, translation studies nowadays familiarize students with the strengths and weaknesses of MT to enable them to post-edit MT output. Hence, AI-based tools have caused a major shift in the training and profession of translators who are by now well-equipped to use MT effectively in their practices (Balling et al., 2014; Chung, 2020). Similar to what happened in the translation classroom, AI-based tools have gained importance in L2 teaching. L2 students are using AI-based tools to translate their ideas from their L1 to their L2, increasingly depending on them to generate written L2 output (Crossley, 2018). Post-editing MT in the L2 writing class has therefore recently emerged as a new research line (Shin & Chon, 2023). Yet, AI-generated text takes writing support to a next level (Gayed et al., 2022, p. 2). L2 students will be tempted to resort to AI-based writing tools to not only *translate*, but directly *generate* ideas in the target language, relying on the machine to 'write' their first draft (or even final version) of a piece they are asked to submit as part of their studies – and by doing so, they might miss important learning opportunities.

It is therefore adamant that L2 writing pedagogy adopts new tasks and instructional concepts to guide students in the critical use of AI-based writing and translation tools in such a way that it benefits the development of their own L2 writing skills.

The current study explores an avenue to integrate ChatGPT in a pedagogically sound manner into the L2 academic writing class, using it as an (imperfect, thus 'coping') model in the revision process. The task at hand is synthesis writing from two sources, a cognitively and linguistically demanding task, even for advanced L2 writers (Solé et al., 2013; Strobl, 2015). At the same time, it is an interesting task for the critical assessment of AI output, since ChatGPT struggles with the balanced selection and integration of relevant information from two sources into one well-structured text, nevertheless creating a stylistically adequate, seemingly cohesive and linguistically accurate output. In L2 writing pedagogy, the question of whether students may use linguistic tools (from spell-checkers and online dictionaries to automatic grammatical feedback) has been discussed extensively and can be divided into a process- versus a product-oriented perspective (Oh, 2022). Adopting a process focus, we acknowledge that in the digital age, such tools form an inevitable and crucial part of L2 writing. At the same time, students need to be prepared to skillfully use these tools and critically interpret their output by developing 'digital literacy', which has been defined as "the ability to properly use and evaluate digital resources, tools and services, and apply it to lifelong learning processes" (Gilster, 1997, p. 220).

In the theoretical background section, we will first zoom in on (L2) writing processes, arguing that outsourcing text generation has a detrimental effect on L2 development. This is followed by a sub-section on the intricacies of the task at hand, synthesis writing, and the pedagogical support that has been shown to foster novice writers' apprenticeship of this task, strategy instruction and modelling. The following subsection is dedicated to beneficial aspects of modelling in L2 writing pedagogy, more specifically to stimulate 'inner feedback' (Nicol, 2021) that can trigger self-revision. We conclude the theoretical background section with suggestions for potential avenues of successful adoption of AI-based writing tools in the L2 writing classroom, zooming in on our own research questions.

Theoretical Background

Writing to Learn (Language): Beneficial Processes Underlying L2 Writing

When outsourcing text generation, there is a risk that many of the processes underlying (L2) writing that support L2 learning disappear. Well-accepted models of writing (Hayes, 1996; Kellogg, 1996) distinguish four subprocesses that cyclically interact with each other. A message starts with its planning phase that involves higher order operations including goal setting and generating ideas retrieved from long-term memory, as well as putting these ideas into a coherent whole, that is, the writing plan. Next, translation processes follow where linguistic encoding takes place, meaning that the planned content is translated into linguistic form drawing on lexical retrieval, morphosyntactic encoding, and application of cohesion. The execution phase is characterized by motor skills when writers type or hand-write the text. The final phase is dedicated to monitoring to check what was written is what was planned and to reread and eventually edit and revise text accordingly. Typically, writers engage in these processes in a dynamic way, that is, monitoring might trigger a new process of planning, translation and execution until the text represents what the writer intended to write. Earlier work has shown that these different phases are more prominent at various stages of the writing process (Révész et al., 2022). For example, at the beginning of text generation, more time is spent on planning; translation and execution dominate the middle stages; monitoring typically increases towards the final stages of the writing process (Khuder & Harwood, 2015; Nicolás-Conesa et al., 2014; Rijlaarsdam & van den Bergh, 1996). Engaging in all these processes of writing supports not only the development and retention of new knowledge (writing to learn) but also supports language learning if the writing takes place in another language than the mother tongue (cf. the concept of "writing to learn language" [Manchón, 2011]). Baaijen and Galbraith's (2018) dual-process model even posits that when starting with unplanned spontaneous text generation, writers develop new ideas and knowledge. Irrespective of their theoretical account, scholars agree that writing is an important activity L2 learners need to perform in the target language if they wish to grow in their L2 competences. Delegating these processes to an AI tool that generates text and therefore content, can therefore be a risk to students' growth.

Not least, outsourcing content planning and organization of ideas might negatively affect critical thinking skills as these are not practiced. Not using cognitive resources on formulation processes to retrieve and select adequate words and structures, not making the effort to generate coherent sentences and cohesive paragraphs is likely to decrease students' competences in this respect, in particular, if the cognitive effort of doing this in an L2 is not used. Even asking a tool to do the monitoring process will confront students with a lack of developing monitoring skills themselves. On the other hand, receiving, reading and further processing correct language that was generated by a tool may also serve as relevant input for L2 writers that can benefit from this to boost their learning.

In short, like in translation studies, students' awareness must be raised about the risks of overly relying on tools that might jeopardize their L2 development as well as their critical thinking and problem-solving skills (Kasneci et al., 2023, p. 7), while also develop the skills of using these tools effectively for their language learning process. In the current study, we explore this in the context of synthesis writing.

Synthesis Writing

Synthesis writing, that is, combining information of two (or more) sources into a new, coherent text, is a particularly challenging task for novice writers, even in tertiary education (Mateos & Solé, 2009). The challenge is partially induced by the fact that it pertains to the domain of knowledge-transforming

writing (Kellogg, 2008), where new knowledge needs to be developed from several inputs. Synthesis writing integrates reading and writing in a recursive process, triggering an internal dialogue that increases opportunities for language development. At the same time, synthesizing information is a core skill underlying academic writing. University students need to develop their academic writing skills irrespective of their study subject (Van Ockenburg et al., 2019).

Earlier work has shown that synthesis writing from multiple sources places three major challenges on the novice (L2) writer. The first challenge is to gain awareness of the *important role of planning and content elaboration* in summarizing processes. Plakans (2009) observed English L2 writers during reading and writing processes in synthesis production. She found that the degree to which the participants employed discourse synthesis processes of organizing, selecting, and connecting during source elaboration had a direct impact on the number of ideas from the source texts they integrated in their syntheses. These findings suggest that students need to develop effective strategies to elaborate the content of the source texts to gain deep understanding (Solé et al., 2013). The second challenge is to understand the *need for linguistic elaboration*. When writing a synthesis, adequate strategies are needed to integrate and rephrase the propositions selected from different source texts. Integrating information from sources, which might all have their own voice, are expected to feed into a new coherent piece of writing that reads fluently. To achieve this goal, novice writers of academic genres need to abandon their natural propensity to focus their attention on linguistic and local aspects. That is, they need to widen their attention from the word and sentence level towards the global text level and work on coherence and cohesion both in terms of content and language (Kellogg, 2008). In an L2 writing context, the tendency to stay close to formulations of source texts is evident given that learners' attention is typically geared towards linguistic accuracy. Finding a 'new voice' for the new text remains challenging also for experienced writers. Thirdly, synthesis writing is highly intertextual, as external textual sources must be incorporated into the writer's own text. Finding an academically acceptable *balance between integration and appropriation of the other's voice* in the writer's own text also requires apprenticeship. Untrained student writers tend to patchwrite when writing from sources by "stitching together elements from one text with elements from another and making some superficial changes to the language" (Pecorari, 2013, pp. 70–71). This strategy often reflects limited elaboration processes and may result in plagiarism. The lack of rephrasing observed in L2 writers is also likely to be related to vocabulary size. Plakans (2009) found that vocabulary size had an impact on the discourse synthesis process of L2 students, both during reading and writing.

Given the potential difficulty of the task, strategy instruction becomes vital for successful task completion. Several studies have highlighted beneficial effects of strategy instruction and modelling for the mastery of complex writing tasks, including synthesis writing (Cumming et al., 2016). Zhang (2013), for instance, found a significant effect for explicit instruction on discourse synthesis strategies combined with in-class writing practice in English L2 writers. In addition to direct strategy instruction, cognitive modelling has been successfully used in interventions to foster advanced writing strategies. Modelling entails learning by observing models that show how to use the strategies in the writing process. For instance, Raedts et al. (2007) used video-based models to support freshmen university students in learning to write academic texts. They found that students who observed coping, and mastery peer models outperformed the ones in the learning-by-doing condition, wrote better post-hoc syntheses in terms of content accuracy and structure, and displayed a more accurate task representation.

To foster students' writing development, different pedagogical approaches have been used over the years. In the following we briefly introduce the role of modelling and (inner) feedback as these guided our own intervention.

Modelling and (Inner) Feedback as a Trigger for Student Induced Text Revision

In L2 writing pedagogy, models have not only been used successfully in supporting the development of writing strategies, but also as a means of corrective feedback (see Garcia Mayo & Labandibar, 2017, for a review). In theory, by comparing their own writing with models, L2 writers are invited to notice the gap between their own interlanguage and target language use (Schmidt, 1990; Schmidt & Frota, 1986). In addition, they will be exposed to writing samples that may function as positive input they could align to and exploit these models to revise their own compositions. During text revision without models or feedback, L2 writers sometimes struggle to notice problems in the first place (Robert et al., 2017). Furthermore, some find it difficult to avoid hyper-revisions (i.e., unnecessary changes that do not change the quality of the text) and over-revision (i.e., changes that introduce errors into the text). Therefore, using models as a means of corrective feedback can help in this process.

In the past decades, modelling has received greater attention from L2 scholars (e.g., Hanaoka, 2007; Kang, 2023; García Mayo & Labandibar, 2017; Roothooft et al., 2022, Wu et al., 2023). Accordingly, writers use the models predominantly to revise vocabulary (e.g., Cánovas Guirao et al., 2015; Hanaoka, 2007; García Mayo & Labandibar, 2017; Wu et al., 2023) and, only to a lesser extent manage to identify content issues. Moreover, lower-level students seem to have rather negative attitudes towards the use of models as writing feedback (García Mayo & Labandibar, 2017; Strobl, 2015). On the other hand, students with positive attitudes managed to benefit from text modelling when engaging in text revision (García Mayo & Labandibar, 2017). Similarly, research with young learners receiving models as corrective feedback showed that proficiency mediated the success of such an approach (Coyle & Roca de Larios, 2020). The most recent findings by Wu et al., (2023) resonate with these suggestions as in their study learners with high language analytic ability benefited more from their intervention than others. This approach also sheds new light on the use of corrective feedback.

A plethora of research has demonstrated that corrective feedback is supportive of language development (e.g., Manchón & Polio, 2022). Still, educators at times question whether students engage enough with the feedback they receive (e.g., Cerezo et al., 2019; Han, 2017, 2019; Santos et al., 2010). In particular, feedback provided in online environments, where novice writers could 'just accept' the suggestions of their corrector (Manchón, 2011), it is questionable how much students indeed learn from corrective feedback if they do not engage with it. Similar concerns are voiced regarding peer feedback (e.g., Fan & Xu, 2020).

More recently, Nicol (2021) has introduced the construct of 'inner feedback'. Accordingly, guided comparison of own text with models can facilitate L2 writing development. This concept relies on the assumption that learners can be steered to create their own 'inner' feedback by asking them to compare their output with good and bad models that have been provided by the teacher. This guided comparison facilitates students' noticing of strengths and weaknesses in their own output. Nicol states that self-generated inner feedback involves students in deeper processing of both content and the language that is needed to express what they intend to write. Consequently, inner feedback has the potential to engage students in those processes that may facilitate long-term retention and learning. In L2 research, studies on feedback processing (e.g., Leow, 2020) come to similar conclusions.

Given that generative AI creates a new reality, where writers will often be tasked (or tempted to give themselves the task) to compare generated text models with their own texts, there is a need to further explore this avenue of research to increase our understanding of how modelling might work in such a context. In addition, effective pedagogies need to be developed that support L2 writers adopting successful strategies to work with generative writing tools.

Writing Pedagogies in Times of Generative AI

Based on the aforementioned research, we identified the following three aspects that L2 educators might include in their teaching to support their students' L2 writing development. Teachers can:

1. come up with tasks that ensure that students engage in the creative and cognitive process of writing independently;
2. help students to make use of tools to support them when translating their intended content into text through processes of linguistic encoding and creating cohesion, while keeping an eye on them still training their linguistic skills in the target language to ensure writing to learn language occurs; and
3. guide students' monitoring and revision processes through cycles of (inner) feedback to support target language development.

In addition, in the context of generative AI, which in theory could take over all the above processes of writing, effective teaching needs to guide students towards a more critical use of AI-models. Acknowledging the limitations of AI-based tools that can generate or translate written text will help students to adopt a critical stance towards the output of these tools, empowering them to effectively use and post-edit AI-generated texts.

Not least, educators must increase their own understanding of "language models in the classroom to ensure that they are being used effectively and not negatively impacting student learning" (Kasneci et al., 2023, p. 7). Practically speaking, students need to learn how to distinguish between aspects of generated text that might serve their learning and aspects that require revision to produce their independent text. This ties in with one of the main educational goals in terms of digital literacy development, according to the European Framework for the Digital Competence of Educators (DigCompEdu) (Redecker, 2017), which is "to use digital technologies to support self-regulated learning processes" (p. 58).

The current study attempts to address some of the challenges that generative AI poses for L2 writing pedagogy. We investigate how AI-based tools such as ChatGPT can be integrated in the L2 writing classroom (Weller, 2023). Specifically, we look at advanced learners of L2 German who engaged in synthesis writing from two sources, a cognitively and linguistically demanding task (Solé et al., 2013). In addition, we investigate the revision process they engaged in after guided comparison of their own drafts with AI-generated syntheses that we used to trigger inner feedback. We were guided by the following research questions:

(RQ1) What do students notice in their own output and in Chat-GPT output based on a guided comparison?
(RQ2) What do students revise in their own texts?

Methods

Participants and Educational Context

Twenty-two students at a Dutch university (14 females, 8 males) aged 18 to 23 years, participated in the present study. All were learners of German as an L2 who were enrolled in German proficiency classes targeting CEFR levels B2 (n = 14) or C1 (n = 8) as part of their BA and MA degrees. Most of the participants (n = 17) had Dutch as their first language and had studied German at high school for at least five years prior to enrolling in university studies. The other five were international students with a variety of first languages and varied schooling experience of German.

Writing and text revision form an inherent part of their program, as tasks often require them to write summaries of what they watched, listened to and discussed in class about the current topic. All participants had prior experience with academic writing, however, in their L2 German, they were mostly acquainted with argumentative writing, having little or no experience with the task of synthesizing the content of two (popular-)scientific texts.

The intervention of the current study was implemented in two proficiency courses and covered four 90-minute sessions. It took place during regular seminars. After ethical clearance, students received full disclosure of the study. Only data of participants that had signed for informed consent were used for further analysis. The other students in the classroom still participated in the pedagogic intervention given that this was part of their regular seminar.

All writing and editing tasks were conducted in Google Docs (campus license), which is the standard environment used for writing and revision assignments in the German proficiency courses that participated in the present study. Thereby, all participants were familiar with the tool as well as its automated correction suggestion function at the time of the data collection. Furthermore, to ensure ecological validity and gain insights into the participants' source use for writing and revision processes, students had full access to all other internet resources. This allowed them to use all auxiliary means they were comfortable with and would resort to in a natural classroom setting, such as online dictionaries and translation software.

Intervention: Instruments, Tasks, and Procedure

Source Texts and ChatGPT Models (GPTM)

To ensure that all participants were familiar with the content of the writing tasks, we chose two topics that had been covered in their current or past proficiency course. Both topics are related to contemporary language variety in German: *Kiezdeutsch* (a variety spoken by multicultural urban youth communities in Germany) and *Anglicisms in the German language*. For each topic, two popular-science articles of similar length (around 450–500 words) were chosen as source texts, with each text presenting different (at times contradicting) facts and arguments on the topic.

Since our short treatment did not involve familiarization and training with ChatGPT as a tool, we provided the participants with AI-generated texts we had prepared for them, including the prompts we used to generate them. In this manner, equal conditions for all participants in the guided comparison and revision sessions were ensured, since they were based on the same models. We generated two GPTM per topic, copy-pasting the two source articles pertaining to the same topic into the dialogue window of ChatGPT (v. 3.5), together with two slightly varying prompts (in German). The first prompt stated: "Please write a synthesis in German of 300–350 words of these two texts. Make sure that the synthesis has an introduction, a main section and a conclusion with a final assessment. The text should be formulated in an academic style." The second prompt was identical with the first, but was enlarged with the sentence: "When paraphrasing, use the subjunctive I."

Workflow and Instructions

The intervention consisted of four entire classroom sessions of 90 minutes each, taking place during two weeks in March 2023. In Sessions (S) 1 and 3, students produced own synthesis of the two texts for each topic, working individually (S1, Topic: *Kiezdeutsch*) and in pairs (S3, Topic: *Anglicisms*), respectively. In S2 and S4, students were tasked to compare their own syntheses written in the prior sessions with two GPTM and to revise their own texts, where deemed necessary, again working

individually (S2) and in pairs (S4) (Figure 1). In the collaborative writing and revision sessions, student pairs were seated next to each other in front of one computer.

Following Cumming et al. (2016), participants received explicit strategy instruction on how to write a good synthesis at two stages of the intervention: (1) Prior to the first writing session, they watched a 10-minute knowledge-clip about the recommended steps for synthesis writing (e.g., first read both texts, highlighting core information) and the requirements of the text form (e.g., balanced representation of both source texts). (2) In S2 and S4, students received a sheet listing eleven quality criteria to critically compare and rate the two GPTM and their own first draft on a 5-point-Likert scale as fulfilling the criteria for good synthesis (Figures 2 and 3). In addition to this quantitative comparative rating, they were tasked to write down a minimum of three strong and three weak points for each of the three texts, i.e., their own synthesis and the two GPTM. After this guided comparison, which typically took about 40 minutes, they used the remainder of the class time to revise their own text, integrating ChatGPT output and/or making any other changes, where deemed appropriate.

For the purpose of data collection targeting the participants' writing and revision behavior as well as the sources they used, all four sessions were recorded with the screen casting software Screenpresso that allows for recording all actions on the screen, including cursor movements and clicks. In addition, the collaborative sessions were audio-recorded using the students' smartphones to capture their conversations during the collaboration process. Since the current article focuses on noticing and revision processes, we only present the analysis of the recordings of the revision sessions (S2 and S4).

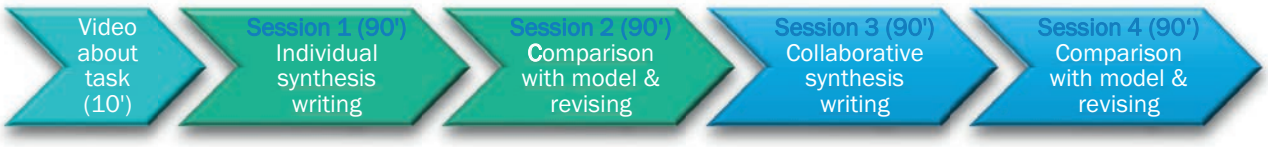

Figure 1 *Workflow of the intervention.*

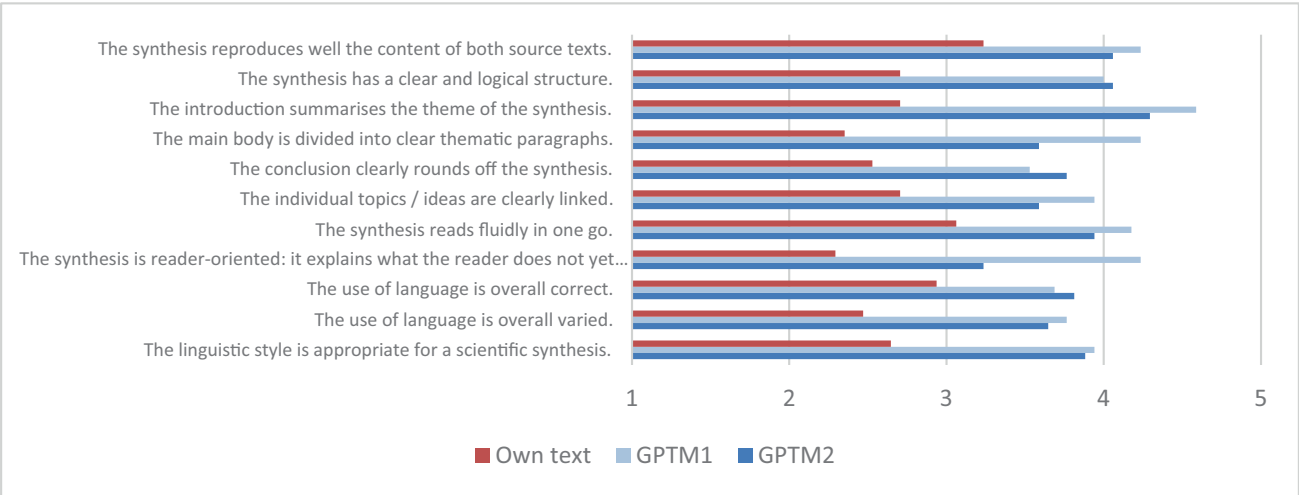

Figure 2 *Text quality assessment by students in Session 2 (individual, n = 17). Mean values between 1 (strongly disagree) and 5 (strongly agree).*

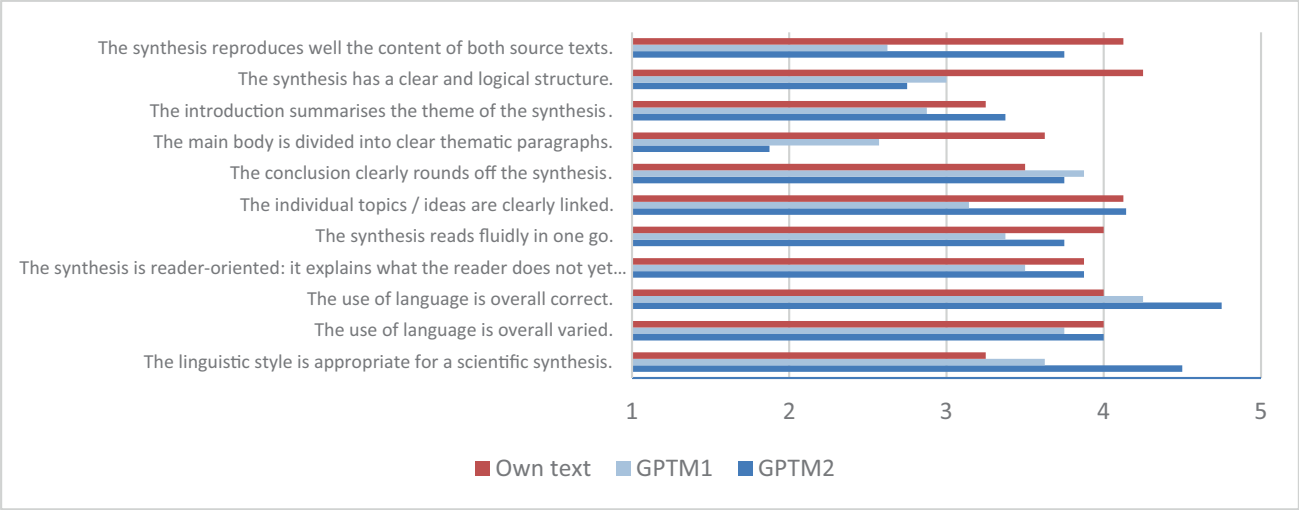

Figure 3 *Text quality assessment by students in Session 4 (8 pairs, n=16). Mean values between 1 (strongly disagree) and 5 (strongly agree).*

Coding and Analysis

To analyze the guided evaluations of the synthesis texts, we calculated the mean values of the 5-point Likert scale ratings of the eleven quality criteria attributed by the students to their own text as well as the two GPTM in S2 and S4. In addition, we grouped their open answers illustrating strong and weak points of all texts according to topic and stance.

For the qualitative coding of students' revision processes, the screen recordings were uploaded in Atlas.ti and coded by three raters according to a codebook. Based on the categories used in previous studies on student revision of translation and writing (Diels, 2022; van Steendam, 2010), we built our own codebook to suit our own specific data. After a first exploratory coding round (2 × 10-minute video excerpts by two coders), the codebook was finetuned, which resulted in a final version that comprises six main categories with several subcategories each (Table 1): revision focus (i.e., the targeted textual feature), necessity, action, success, trigger (if identifiable, e.g. the GPTM or automated correction suggestions by Google) and information sources that were consulted in the revision process including the trigger observed. Several subcodes could apply to one revision episode.

Under the code revision focus, we grouped typography, orthography, morphology (e.g., case marking) and interpunction into the subcode 'local' because they are frequently detected by the automated correction system of Google Docs, providing correction suggestions that can be accepted with one mouse click. 'Grammar revisions' on the other hand refer to changes transcending the word level, such as, active/passive transformations.

For further analyses, we identified six participants based on completeness of data set (cf. Section **Limitations and Directions for Further Research**). The six focus participants form a representative sample of the entire cohort regarding language level in L2 German (four participants were at level B2, two at C1), L1 (one Spanish, one Polish and four Dutch native speakers) as well as gender (one male and five female students). The final corpus coded for the analyses presented in the current paper comprises nine videos: six videos of the individual session (S2) and three of the pair revision session (S4). The videos were coded by three coders, the first author and two linguistically trained research assistants. To ensure inter-coder reliability, 2 × 10-minute videos were double-coded. Inter-coder agreement computed in Atlas.ti was higher than 0.8 (Krippendorf's alpha) for all categories.

During the remaining coding process, the three coders met three times, discussing selected episodes and corner cases to ensure maximum convergence in coding behavior.

Results

First, we report on the students' quality assessment of the synthesis texts (i.e., their own text and the two GPTM) to illustrate their noticing of relevant quality issues based on the assessment sheet. Second, we present the results of the coding analysis of their revision behavior in terms of frequency of revision per (sub)categories, also grouping results of a co-occurrence analysis of categories. We only present descriptive statistics of our data, since the small number of participants does not allow for sound inferential statistics.

Quality Assessment of ChatGPT Models (GPTM) and Own Drafts by Students

Figures 1 and 2 show the mean values of the ratings of students' own text (draft 1 before revision) in comparison with the two GPTM in the individual and collaborative judgment and revision sessions, respectively. The student ratings reveal that they thought their own texts (draft 1 before revision) were of lower quality than the GPTM in the individual session (Figure 1) for all eleven criteria: mean scores range from 2.3 (for *clear thematic division of main body)* to 3.2 (for *scientific writing style*) for own texts; and from 3.2 (for *clear thematic division of main body* in GPTM2) to 4.6 (for *correct language use* in GPTM1) for GPTM. Interestingly, in the collaborative session (Figure 2), students rated their own texts markedly higher with mean scores ranging from 3.2 (for *scientific writing style* and *introduction*) to 4.5 (for *clear and logical structure),* positioning their own text in between the two GPTM for seven criteria and even highest for the four indicators *content representation, clear and logical structure, thematic division of main body* and *fluent reading experience.* Ratings for GPTM in the collaborative session 4 had lower mean values starting at 1.9 (given to *thematic division of main body),* while the acknowledgement of *correct language use* remains stably high for GPTM (max. 4.8).

For the qualitative assessment answering the request to pinpoint strong and weak points of each text, we analyzed the responses provided by our six focus participants. These responses are in line with the quantitative results of the group assessment of the GPTM. Our six focus participants positively commented on structure and correct, academic language use (e.g., good use of cohesive devices to structure the text), the varied and broad lexicon, and the integration of specific content elements. The lower ratings given in session 4 tie in with a notable increase of weak points brought forward. For example, the lack of originality and creativity resulted in a 'dull' reading experience as exemplified by the following comment: "Given the topic of *Kiezdeutsch* as a highly creative language variety, it is a pity that ChatGPT itself does not use creative language." (Students' quotes were translated from original German by the first author.) Moreover, the integration of 'invented' information that was not present in the source texts made it untrustworthy, according to the students' comments. In other words, students discovered that ChatGPT tends to 'hallucinate'. Lastly, but importantly, the low degree of integration of both source texts made that "the text reads as two summaries, but not as a synthesis of the two texts."

Revision of Own Texts Based on GPTM

Table 1 provides an overview of the revision behavior observed for the six focus participants in all nine recordings. The first column states the names of the six code groups and the associated subcodes, which are ordered according to overall frequency (cf., second column). The third and fourth column show the average distribution of the subcodes per code group, that is, the mean of coded episodes in

Table 1 *Frequency of revision behavior per (sub)category as percentage of coded episodes for all sessions (n = 9), individual session 2 (n = 6) and the collaborative session 4 (n = 3), respectively*

	All	S2 (mean)	S4 (mean)
Revision focus	n = 230	n = 28	n = 20
● content	31%	32%	28%
● local (word-internal and interpunction)	27%	29%	20%
● lexical choice	14%	12%	20%
● structure	8%	9%	5%
● cohesion	7%	7%	8%
● other (layout, word count)	7%	8%	5%
● grammar (word-external)	6%	4%	13%
Revision necessity	n = 230	n = 28	n = 20
● unnecessary	54%	52%	60%
● necessary	46%	48%	40%
Revision success	n = 244	n = 28	n = 20
● improvement	63%	57%	79%
● neutral	23%	27%	11%
● aggravation	14%	16%	10%
Revision action	n = 230	n = 28	n = 20
● substitution	40%	46%	22%
● insertion	34%	31%	45%
● deletion	14%	13%	17%
● no action	10%	8%	13%
● move	2%	2%	3%
Revision trigger	n = 230	n = 28	n = 20
● not identifiable	50%	54%	38%
● Google suggestion	29%	34%	15%
● peer discussion	11%	0%	43%
● ChatGPT model	10%	12%	3%
● source texts	0,4%	0,6%	0%
Information sources	n = 242	n = 28	n = 20
● not identifiable	39%	38%	42%
● Google suggestion	31%	35%	18%
● ChatGPT model	14%	15%	10%
● other online tools	6%	7%	5%
● peer discussion	5%	0%	21%
● Google translate	2%	2%	0%
● other	2%	1%	5%
● Google search	1%	2%	0%

six recordings of the individual session 2 and the mean of the coded episodes in the three recordings of the collaborative session 4. Accordingly, slightly more revision episodes took place in the individual (n = 28) than collaborative sessions (n = 20).

Overall, a clear predominance of revision focus on content and local issues is visible (both about a third), with about a sixth of revisions pertaining to lexical choice. In both revision sessions, more than half of the revisions were unnecessary (i.e., hypercorrections). Still, 65% of all revisions were successful, amounting up to 86% in the collaborative context that led to improvement of the text. Regarding revision action, it is interesting to note that in most cases where a revision attempt was identifiable (i.e., the cursor hovered or mouse clicks occurred over a specific text element), revisions were actually made: 'no action' was coded only in 6% (S2) and 9% (S4) of the identifiable revision attempts. Frequencies of revision actions show that individuals tended to substitute text while pairs inserted information (about half of the time) followed by deletions.

Table 2 *Co-occurrence analysis of revision focus versus revision trigger, revision success and revision necessity*

Focus	**Content n = 71**	**Cohesion n = 17**	**Lexical Choice n = 32**	**Local n = 62**	**Structure n = 18**	**Grammar n =1 4**	**Other n = 16**
Trigger							
ChatGPT model n = 22	21	1					
Google suggestion n = 66	2	1	4	48	4	6	1
Not identifiable n = 115	36	13	21	11	13	6	15
Peer discussion n = 26	11	2	7	3	1	2	
Source texts n = 1	1						
Success							
Improvement n = 153	52	14	21	48	6	11	1
Neutral n = 56	8	3	7	11	10	2	15
Aggravation n = 35	18	4	4	6	2	1	
Necessity							
Necessary revision n = 106	20	6	16	50	4	8	2
Unnecessary revision n = 124	51	11	16	12	14	6	14

Note: On some occasions, revisions can lead to overall improvement (e.g., adding a relevant content aspect), while at the same time introducing new errors (e.g., grammar problem). Therefore, some focus episodes received two codes for 'revision success'. This explains why the sum of subcodes and the overall numbers per revision focus do not always add up.

Table 3 *Co-occurrence analysis of revision necessity vs. revision success*

Necessity	Necessary Revision n = 106	Unnecessary Revision n = 124
Success		
Improvement n = 153	94	59
Neutral n = 56	8	48
Aggravation n = 35	9	26

Note: Some focus episodes received two codes for 'revision success'. This explains why the sum of subcodes and the overall numbers per revision focus do not always add up.

Revisions were mainly triggered by peer discussions (in collaborative settings, as established by the audio recordings), Google Docs suggestions, and around 10% by the GPTM. Still, the number of unidentifiable revision triggers is high in both sessions, covering 47% (S2) and 35% (S4) of the cases. Finally, yet importantly, for their revisions, students rarely made use of information sources other than Google suggestions, GPTM or their peers (in session 4). That is, Google search actions and online dictionaries (e.g., to check lexical meaning or collocations) were almost never consulted.

Co-occurrence Analysis

To identify high-frequent co-occurrences between revision focus, revision trigger and revision success, we ran these analyses in Atlas.ti. Tables 2 and 3 present the number of episodes in which the two codes at hand co-occurred. Empty cells mean that no co-occurrences took place in this category. Shades of green represent frequent co-occurrences, with darker colours indicating higher frequencies.

As can be seen from Table 2, half of the **content revisions** were inspired by an unidentifiable trigger, followed by GPTM as the most frequent identifiable trigger in over 30% of the cases. Moreover, although content revisions were unnecessary in two thirds of the cases, more than two thirds of content revisions led to an improvement of the texts. On the contrary, two thirds of the **local revisions** triggered by Google Docs' correction suggestions were indeed necessary and induced an improvement. Google Docs also was an important trigger for the small category of **grammar revisions**, of which 60% were deemed as necessary, and led to improvement in four out of five times. The trigger for **lexical choice revisions** (the third most frequent category), was not identifiable in two thirds of the cases; although a revision only was coded as necessary in half of the cases, the changes made, led to textual improvement in a third of all the cases. **Structure** (e.g., re-ordering information and subdividing paragraphs) and **cohesion** (e.g., adding linking words between sentences and paragraphs) were targeted less frequently and to an equal extent. Revisions of both categories were deemed unnecessary by the coders in two thirds of the cases, however, cohesion revision led to improvement in four out of five cases, whereas structural changes did so only for 45%. Finally, yet importantly, the co-occurrences of unnecessary revisions that led to aggravations in textual quality were rather rare, since they only concern about a fifth of all unnecessary revisions (Table 3).

Discussion

This study set out to test the value of a classroom-based intervention in a university course for German as an L2 where we guided students to adopt ChatGPT as a writing buddy during individual and collaborative synthesis writing, inspired by post-editing practices in translation pedagogy and the construct of 'inner' feedback introduced by Nicol (2021). Accordingly, students received synthesis

models generated by generative AI and were asked to rate and compare their own text with the generated models, followed by revision of their own texts. We analyzed the students' ratings of all syntheses and performed an in-depth analysis of the revision processes six focus participants engaged in.

Since noticing is a prerequisite for revision, our **first research question** asked what students would notice in Chat-GPT output and in their own texts, based on the guided comparison. Following Kasneci et al. (2023), we hypothesized that they would focus on issues related to grammatical accuracy, vocabulary use and writing style in their own texts, while they might notice content bias and flaws in ChatGPT output. The quantitative ratings of the participants as well as their qualitative descriptions of strengths and weaknesses indeed partly tie in with these hypotheses. Students rated their own output consistently lower in terms of correct language use and appropriate writing style in comparison with the ChatGPT models (GPTM). However, they also criticized the GPTM for a lack of creative language use, especially in the first, individual revision session (S2). This suggests that our participants might not yet have been fully acquainted with the task requirements of synthesis writing at the outset of the intervention, therefore transferring their prior experience with **argumentative** academic writing. Although they made some critical remarks about content selection in GPTM and its trustworthiness, they still rated GPTM high in terms of content, especially in S2.

This explanation ties in with the markedly higher ratings of students' own drafts in comparison with the two GPTM in the second, collaborative revision session (S4) as opposed to the first, individual session (S2). The higher ratings of their own drafts also correspond with the fact that on average, more revisions were made in the individual (S2) than in the collaborative (S4) revision session. The two possible explanations for this seemingly increased confidence with their own text production are task effect and effect of the setting, respectively. Students might have felt in S4 that they learnt from S2 and/or they might feel more confident about collaborative (S3) than individual text production (S1) and therefore engaged less in revisions as a pair (S4). In any case, the guided comparison showed its effectiveness in engaging the students in a critical stance towards their own texts as well as texts generated by AI tools, hopefully preventing them from using such tools blindly.

Our **second research question** asked what students revised in their own texts after the guided comparison with GPTM. Based on previous studies of revision processes informed by models (Cánovas Guirao et al., 2015; Hanaoka, 2007; García Mayo & Labandibar, 2017; Roothooft et al., 2022), we hypothesized that local problems, such as grammar and lexical choice, would be tackled more frequently than global textual concerns, such as content and structure. This hypothesis was not confirmed by our results. A clear majority of revision actions concerned content issues, while structure and cohesion, two important aspects of higher-order concerns, were targeted to a far lesser extent. This is especially notable given the frequent positive mention of structure and fluent reading of GPTM in the guided judgment phase prior to revision.

On the contrary, the hypothesized correlation between focus and trigger is clearly observable in our data: the foremost identifiable trigger for content revisions was GPTM, whereas the great majority of local problems were (correctly) solved using Google Docs automated correction suggestions. In combination with the low number of consultations of other information sources, it can be deduced that Google Docs suggestions were followed rather blindly, which turns out to be an efficient and effective choice.

Importantly, we expected to observe occurrences of hyper-revisions and over-revisions. Our data show that more than half of the revisions made were coded as unnecessary and would therefore be classified

as hyper-revisions, which are deemed inefficient (Robert et al., 2017). However, co-occurrence analyses revealed that even unnecessary changes (e.g., adding content or rephrasing an error-free textual fragment) in half of the cases ultimately led to improvement of the text, which means that they were effective revisions. Only in 22% of the cases, unnecessary revisions were co-coded as aggravations; therefore, over-revision did not seem to be an important problem.

In sum, the clear co-occurrence figures between revision focus and trigger together with the high number of successful revisions show that students seemed to be well equipped to revise their texts, while skillfully drawing on the different sources they had at their disposal. On a cautionary note, we would like to pinpoint that our participants were advanced students of the target L2 which has been described as an important factor for effectiveness of models in the revision process (García Mayo & Labandibar, 2017, p. 110; Wu et al., 2023).

Finally, yet crucially, it is worth contemplating the high number of unidentifiable revision triggers in both sessions (half and a third of all revision episodes, respectively). In these episodes, changes were applied in the course of re-reading without identifiable consultation of the GPTM or any other information source. A possible explanation is that the source texts were provided both electronically and on paper, rendering re-reading of the source texts cumbersome to identify for coding. Yet, the coders checked both source texts and GPTM for potential textual borrowing and coded them as potential triggers accordingly, whenever changes of content and larger text chunks occurred. Therefore, it is safe to assume that the actual trigger in these cases is noticing gaps in the students' own language output (Schmidt, 1990; Schmidt & Frota, 1986) while re-reading. This is even more plausible given the participants' advanced language proficiency (García Mayo & Labandibar, 2017, p. 110) and the task sequence of writing – guided judgement task based on models – revising that has been documented to promote 'inner feedback' (Nicol, 2021).

Limitations and Directions for Future Research

This study is not without limitations. Given the ecologically valid context of implementing this study as an intervention in a real classroom, we experienced substantial data loss. While initially more than 20 students had signed up to participate, absences during seminar sessions that are typical for a university course caused a much smaller number of complete data sets. We compensated the small size by an in-depth analysis of the revision behaviors of our six focus participants. Still, a larger data set would allow for statistical inference testing, thus making it possible to be more confident about the transferability of our findings.

Another limitation is the task effect inherent in an intervention design including task repetition, which we could not mitigate by counterbalancing due to a lack of time and resources. This means that over the course of the four sessions, students became more acquainted with effective strategies for synthesis writing and revision, gaining not only confidence, but also competence in the task at hand. Future research adopting our or similar designs, should consider mitigating these effects by providing more opportunities for informed practice, including extended strategy instruction, before the start of the intervention. Ideally, future interventions should also include process-based modelling (Raedts et al., 2007) in addition to using AI-generated models to stimulate revision.

In this study, we have only looked at the task of synthesis writing, but it is to be expected that other writing genres (e.g., abstract writing, argumentative writing, narrative texts) lend themselves to working with AI-generated models and promote inner feedback. Expanding our methodology to these types of texts, would likely further our understanding of what role ChatGPT and other tools can play in future writing pedagogy.

Conclusion and Pedagogical Implications

This study set out to investigate the potential of ChatGPT output as a model in the revision of a challenging academic writing task, L2 synthesis writing, targeting both noticing and actual revision behavior. Noticing was stimulated by a guided judgment task before the actual revision. The combined writing and revision task was carried out twice, once individually and once in collaborative pairs.

The collected data include quantitative and qualitative evaluations of students' own text in comparison with two ChatGPT models provided in each revision session and screen- and audio-recordings of the revision sessions that were coded for revision episodes along six main categories, including focus, trigger and success of revisions.

Our results showed that the guided judgment tasks phase clearly led to high engagement in the following revision, triggering noticing of content and language flaws in the L2 students' own written output. The noticing was partly triggered by the models and Google correction suggestions, but also for a large part based on own insights without identifiable triggers, showcasing successful practices of 'inner feedback'. Furthermore, we found that students relied to a great extent (blindly) on Google correction suggestions for local problems, such as orthographic errors and incorrect case/gender marking and resorted to their peer (in case of collaborative revision) and/or the ChatGPT models for content revisions. Content was by far the focus of the revisions, contrary to findings of earlier studies on model-based revision that reported a main revision focus on lexical choice. Higher-order concerns other than content, such as structure and cohesion, were targeted to a far lesser extent in the revision, even if the superiority of the models in that regard had been noticed in the evaluation phase. In conclusion, students made effective choices between the different means of support that were available.

Our analyses further revealed that collaboration proves to be an important trigger for model-based revision, with nearly half of all revision episodes in the collaborative sessions being sparked by peer discussions. Those discussions most often focused on content issues, followed by lexical choices. We therefore recommend to further explore the role of GPTM in collaborative L2 writing and revision.

To conclude, the results of this intervention using ChatGPT-output as models in a writing – judgment – revision task sequence provide evidence that AI tools based on large language models can promote writing development in the advanced L2 classroom. Moreover, through critical engagement with the models, students' digital literacy skills can be nurtured. By using AI-generated texts as models and guiding them through a critical comparison, we tapped into learner´s digital competences that need to be facilitated by educators, according to the European DigCompEdu framework (Redecker, 2017). According to this framework, students need to be taught "to manage risks and use digital technologies safely and responsibly" (p. 84). Moreover, by informing students about the prompts used to create these models, we empowered them to use the technology for their own, self-directed learning, and thus, "[t]o use digital technologies in innovative ways to create knowledge" (p. 86).

As such, we have been successful in developing a pedagogical intervention that fulfilled the roles we envisioned: engaging and guiding students through writing and monitoring processes in the L2 that will eventually benefit their L2 learning.

Acknowledgements

We thank Maarten Pol and Larissa Weber for their support as student assistants in this project. We thank all our participants and students for engaging in the activities and sharing their insights.

References

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

Balling, L. W., Carl, M., & O'Brian, S. (Eds.). (2014). *Post-editing of machine translation: Processes and applications*. Cambridge Scholars Publishing.

Cánovas Guirao, J., Roca de Larios, J., & Coyle, Y. (2015). The use of models as a written feedback technique with young EFL learners. *System, 52*, 63–77. https://doi.org/10.1016/j.system.2015.04.002

Cerezo, L., Manchón, R. M., & Nicolás-Conesa, F. (2019). What do learners notice while processing written corrective feedback? A look at depth of processing via written languaging. In R. P. Leow (Ed.), *The Routledge handbook of second language research in classroom learning* (pp. 171–185). Routledge. https://doi.org/10.4324/9781315165080

Chung, E. S. (2020). The effect of L2 proficiency on post-editing machine translated texts. *Journal of Asia TEFL, 17*(1), 182. http://dx.doi.org/10.18823/asiatefl.2020.17.1.11.182

Coyle, Y., & de Larios, J. R. (2020). Exploring young learners' engagement with models as a written corrective technique in EFL and CLIL settings. *System, 95*, 102374. https://doi.org/10.1016/j.system.2020.102374

Crossley, S. A. (2018). Technological disruption in foreign language teaching: The rise of simultaneous machine translation. *Language Teaching, 51*(4), 541–552. https://doi.org/10.1017/S0261444818000253

Cumming, A., Lai, C., & Cho, H. (2016). Students' writing from sources for academic purposes: A synthesis of recent research. *Journal of English for Academic Purposes, 23*, 47–58. https://doi.org/10.1016/j.jeap.2016.06.002

Diels, E. (2022). *The effects of corpus-focused instruction on sociolinguistic competence: A mixed-methods study into stylistic translation revision competence in English*. Unpublished PhD dissertation. University of Antwerp. URL: https://repository.uantwerpen.be/docstore/d:irua:15131

Fan, Y., & Xu, J. (2020). Exploring student engagement with peer feedback on L2 writing. *Journal of Second Language Writing, 50*, 100775. https://doi.org/10.1016/j.jslw.2020.100775

García Mayo, M. P., & Labandibar, U. L. (2017). The use of models as written corrective feedback in English as a foreign language (EFL) writing. *Annual Review of Applied Linguistics, 37*, 110–127. https://doi.org/10.1017/S0267190517000071

Gayed, J. M., Jonson Carlon, M. K., Oriola, A.M., & Cross, J.S. (2022). Exploring an AI-based writing assistant's impact on English language learners. *Computers and Education: Artificial Intelligence, 3*. https://doi.org/10.1016/j.caeai.2022.100055

Gilster, P. (1997). *Digital literacy*. Wiley.

Han, Y. (2017). Mediating and being mediated: Learner beliefs and learner engagement with written corrective feedback. *System, 69*, 133–142. https://doi.org/10.1016/j.system.2017.07.003

Han, Y. (2019). Written corrective feedback from an ecological perspective: The interaction between the context and individual students. *System, 80*, 288–303. https://doi.org/10.1016/j.system.2018.12.009

Hanaoka, O. (2007). Output, noticing, and learning: An investigation into the role of spontaneous attention to form in a four-stage writing task. *Language Teaching Research, 11*(4), 73–93. https://doi.org/10.1177/1362168807080963

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

Kang, E. Y. (2023). Model-based feedback for L2 writing revision: The role of vocabulary size and language aptitude. *International Journal of Applied Linguistics*, 1–14. https://doi.org/10.1111/ijal.12480

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kellogg, R. (1996). A model of working memory in writing. In C. M. Levy, & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–72). Lawrence Erlbaum.

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

Khuder, B., & Harwood, N. (2015). L2 writing in test and non-test situations: Process and product. *Journal of Writing Research, 6*, 233–278. https://doi.org/10.17239/jowr-2015.06.03.2

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing, 34*(3), 321–336. https://doi.org/10.1177/0265532216663991

Leow, R. P. (2020). L2 writing-to-learn: Theory, research, and a curricular approach. In R. M. Manchón (Ed.), *Writing and language learning: Advancing research agendas* (pp. 95–120). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.56

Manchón, R. M. (Ed.) (2011). *Learning-to-Write and Writing-to-Learn in an additional language*. John Benjamins Publishing Company. https://doi.org/10.1075/lllt.31

Manchón, R. M., & Polio, C. (Eds.) (2022). *The Routledge handbook of second language acquisition and writing*. Routledge.

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

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756–778. https://doi.org/10.1080/02602938.2020.1823314

Nicolás-Conesa, F., Roca de Larios, J., & Coyle, Y. (2014). Development of EFL students' mental models of writing and their effects on performance. *Journal of Second Language Writing, 24*, 1–19. https://doi.org/10.1016/j.jslw.2014.02.004

Oh, S. (2022). The use of spelling and reference tools in second language writing: Their impact on students' writing performance and process. *Journal of Second Language Writing 57*, 1–12. https://doi.org/10.1016/j.jslw.2022.100916

Pecorari, D. (2013). *Teaching to avoid plagiarism. How to promote good source use*. Open University Press.

Plakans, L. (2009). The role of reading strategies in integrated L2 writing tasks. *Journal of English for Academic Purposes, 8*(4), 252–266. https://doi.org/10.1016/j.jeap.2009.05.001

Raedts, M., Rijlaarsdam, G., van Waes, L., & Daems, F. (2007). Observational learning through video-based models: Impact on students' accuracy of self-efficacy beliefs, task knowledge and writing performances. In S. Hidi & P. Boscolo (Eds.), *Writing and motivation*, (pp. 219–238). Springer.

Redecker, C. (2017). *European Framework for the Digital Competence of Educators: DigCompEdu*. Punie, Y. (ed). *EUR 28775 EN*. Publications Office of the European Union. doi:10.2760/159770, JRC107466

Révész, A., Michel, M., Lu, X., Kourtali, N., Lee, M., & Borges, L. (2022). The relationship of proficiency to speed fluency, pausing, and eye-gaze behaviours in L2 writing. *Journal of Second Language Writing, 58*, 100927. https://doi.org/10.1016/j.jslw.2022.100927

Rijlaarsdam, G., & van den Bergh, G. (1996). The dynamic of composing—An agenda for research into an interactive compensatory model of writing: Many questions, some answers. In C. M. Levy, & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 107–126). Lawrence Erlbaum.

Robert, I., Remael, A., & Ureel, J. J. (2017). Towards a model of translation revision competence. *The Interpreter and Translator trainer, 11*(1), 1–19. https://doi.org/10.1080/1750399X.2016.1198183

Roothooft, H., Lázaro-Ibarrola, A., & Bulté, B. (2022). Task repetition and corrective feedback via models and direct corrections among young EFL writers: Draft quality and task motivation. *Language Teaching Research*, 136216882210820. https://doi.org/10.1177/13621688221082041

Santos, M., López Serrano, S., & Manchón, R. M. (2010). The differential effect of two types of direct written corrective feedback on noticing and uptake: Reformulation vs. error correction. *International Journal of English Studies, 10*(1), 131–154. https://doi.org/10.6018/ijes/2010/1/114011

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*, 129–158. https://doi.org/10.1093/applin/11.2.129

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language. A case study of an adult learner of Portuguese. In R. Day (Ed.), *Talking to learn: Conversation in second language acquisition*, (pp. 237–326). Newbury House.

Shin, D., & Chon, Y. V. (2023). Second language learners´ post-editing strategies for machine translation errors. *Language Learning & Technology, 27*(1), 1–25. https://hdl.handle.net/10125/73523

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

Strobl, C. (2015). Attitudes towards online feedback on writing: Why students mistrust the learning potential of models. *ReCALL, 27*(3), 340–357. https://doi.org/10.1017/S0958344015000099

Van Ockenburg, L., van Weijen, D., & Rijlaarsdam, G. (2019). Learning to write synthesis texts: A review of intervention studies. *Journal of Writing Research, 10*(3), 402–408. https://doi.org/10.17239/jowr-2019.10.03.01

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Bergh, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction, 20*(4), 316–327. https://doi.org/10.1016/j.learninstruc.2009.08.009

Weller, D. (2023). *ChatGPT for language teachers: The ultimate prompt handbook for AI productivity*. Stone Arrow Publishing.

Wu, Z., Qie, J., & Wang, X. (2023). Using model texts as a type of feedback in EFL writing. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1156553

Zhang, C. (2013). Effect of instruction on ESL students' synthesis writing. *Journal of Second Language Writing, 22*(1), 51–67. https://doi.org/10.1016/j.jslw.2012.12.001