Journal of Second Language Writing 67 (2025) 101185

Contents lists available at ScienceDirect

Journal of Second Language Writing

journal homepage: www.elsevier.com/locate/jslw

Collaborative writing based on generative AI models: Revision and deliberation processes in German as a foreign language

Marije Michel[a,*,1], Iryna Bazhutkina[a,b,2], Niklas Abel[a,c,d,3], Carola Strobl[b,4]

[a] University of Groningen, Center for Language and Cognition Groningen, the Netherlands
[b] University of Antwerp, Belgium
[c] Amsterdam University of Applied Sciences, the Netherlands
[d] University of Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
ChatGPT
Collaborative writing
L2 writing pedagogy
Functional adequacy

ABSTRACT

The introduction of generative AI (GenAI, e.g., ChatGPT) has transformed second language (L2) writing practices. Writers need to learn how to critically employ the tools' affordances to maximise their learning and L2 instructors face the challenge of guiding students through these novel contexts (cf. Sasaki, 2023). Also collaborative writing (CW) interventions have to find ways how to include GenAI as another technology that impacts the effectiveness of CW as an instructional L2 writing tool (cf. Storch, 2022; Su et al., 2023). In this study, we present an in-depth analysis of how GenAI influences revision and deliberation processes during CW within four writing pairs (N = 8; aged 18–23) engaging in a classroom-based intervention that tasked them to compare their own writing with GenAI models. We analysed screen recordings of evolving revision behaviours, changes in text quality as well as joint discussions among pairs. Complementing Strobl, Menke-Bazhutkina, Abel, and Michel (2024), the current paper (i) relates writers' revision behaviour to text quality evaluated for functional adequacy (Kuiken & Vedder, 2017); and (ii) presents a fine-grained review of dialogues to identify patterns of CW revisions in interaction with GenAI models. Findings shed new light on CW models (Storch, 2009) in the digital era where GenAI is implemented in L2 writing and revision instruction.

1. Introduction

In recent years, second language (L2) writing instruction has seen a growing number of tools transforming writing practices such as early on spell-checkers and translation tools (Oh, 2022). The recent introduction of generative artificial intelligence (GenAI) based on large language models (LLMs) like ChatGPT, forms not just the newest, but probably the most radical addition to technology-enhanced

[*] Corresponding author.
E-mail addresses: m.c.michel@rug.nl (M. Michel), i.a.menke-bazhutkina@rug.nl (I. Bazhutkina), n.abel@rug.nl (N. Abel), carola.strobl@uantwerpen.be (C. Strobl).
[1] https://orcid.org/0000-0003-1426-4771
[2] https://orcid.org/0009-0003-6105-0423
[3] https://orcid.org/0009-0009-1158-9555
[4] https://orcid.org/0000-0002-6306-3514

https://doi.org/10.1016/j.jslw.2025.101185
Received 1 June 2024; Received in revised form 20 January 2025; Accepted 31 January 2025
Available online 7 March 2025

writing. In order to maximise their learning, current generations of L2 writers need to learn to critically assess and leverage these technologies. Similarly, L2 instruction faces the challenge of guiding students through the novel affordances of GenAI (cf. Sasaki, 2023; Warschauer et al., 2023). These technological advancements have also introduced new (digital) forms of collaborative writing (CW) and literacy practices, which have sparked a productive line of L2 research investigating its effectiveness (cf. Storch, 2019) for both learning to write and writing to learn (Manchón, 2011). Accordingly, CW supports L2 development because deliberations, that is, weighing in arguments and counterarguments to come to a shared conclusion (Casado-Ledesma et al., 2021), on how to express joint ideas in a shared text focuses a writers' attention to language form, which in turn promotes noticing (Storch, 2002, Schmidt, 1990). Yet, Storch's (2009) work has shown that collaboration is only fruitful if pairs indeed collaborate – in other words, the pairing of individuals with their specific characteristics has an impact on how the shared efforts might result in a successful joint text. With the availability of GenAI, yet another player can be added to the interaction between L2 writing peers and the evolving text, that is, for example, an AI-generated model (Su et al., 2023) or AI-based feedback (Wiboolyasarin et al., 2024). Which role GenAI can play in L2 writing, how it can be productively used by educators to support L2 learning, and what effects the implementation of GenAI will have on L2 development, still has to be explored. In the present study, we strive to contribute to the advancement of research on the use of GenAI for L2 teaching and learning with a focus on collaborative revision processes. We present data of pairs of students of L2 German at a Dutch university who work together on revising a synthesis text they wrote using CW based on a guided comparison with GenAI-created model syntheses.

To situate our study in the field, we will first review recent literature on collaborative writing and revision before highlighting some of the work on GenAI-based L2 writing and elaborating on our own research in this area.

2. Literature review

Long before Manchón (2011) introduced the distinction between Learning to Write (L2W) and Writing to Learn (W2L – or W2LL if it is about learning the language), it was clear that writing forms an important locus for L2 learning (Cumming, 2020). The permanence of written text and increased time that is provided during text generation allows L2 learners to focus on language form at all phases of the writing process, starting with pondering about the content and organisation of a text during planning, the translation of these ideas into words and sentences during formulation, and the monitoring of the language produced and coherence of the text as a whole during revision processes (Kormos, 2023). Other than ephemeral oral production, where everything said disappears within seconds, the permanence of writing allows for heightened attention to language form, and accordingly creates a unique opportunity for L2 development. These benefits concern both expanding L2 knowledge, for example, because learners can look up more advanced words and expressions, as well as strengthening accurate Form-Use-Meaning-Mappings (FUMMs, Verspoor, 2017). Hence, during writing L2 learners can take their time to, for example, apply explicit knowledge about correct grammatical forms and adequate means to sign-post coherence. As advocated by Storch (2002), the L2 learning potential of writing activities is increased in collaborative settings. In the following, our literature review first discusses findings from early work on CW and revision on the impact on L2 development. Second, given our interest in the L2 learning potential of collaborative revision in combination with GenAI, we provide an overview of recent research on the impact of Web2.0 and GenAI on L2 writing processes and pedagogies. Finally, the literature review presents findings of our own intervention study (Strobl et al., 2024) concerning the impact of individual vs. collaborative settings on revision behaviour, which laid the grounds to zoom in on the research question about the role of collaborative patterns tackled in the present article.

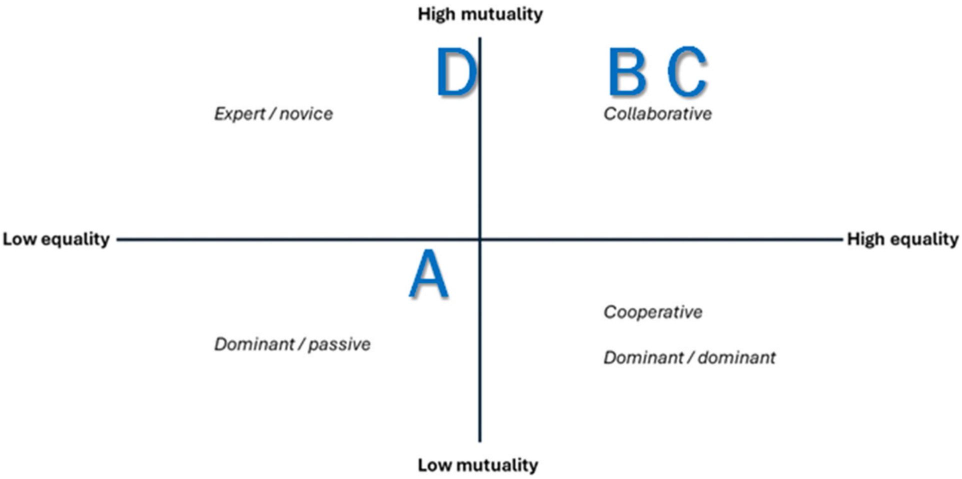

Fig. 1. *Collaborative revision behaviour of four focus pairs in* Storch's (2013) *model.*

2.1. Collaborative writing and revision

Over the past decades, Storch (e.g., 2002, 2009, 2018, 2022) has been a driving force researching the value of CW, that is, when two (or more) writers share the ownership and responsibility for a text that is planned, written and revised by joint efforts. During CW, pairs of learners discuss the processes of planning, formulation and revision. This triggers what Swain (2006) coined languaging, that is "the process of making meaning and shaping knowledge and experience through language" (p. 98). Accordingly, CW promotes Language Related Episodes (LREs, Swain and Lapkin, 1998) and increases noticing (Schmidt, 1990) at all levels of text production (i.e., content and language) as part of pushed output (Swain, 1985). CW forms a context of collective scaffolding (Donato, 1994), where the "availability of immediate assistance from peers to address holes in knowledge learners discover in the process of writing" allows L2 writers to "pool their linguistic knowledge" (Storch, 2022, p. 27). LREs also occur in collaborative revision, taking the form of negotiations that are evaluative (intended to fix errors), social (between-writers dialogue) or procedural (task-related discussions) in nature (Hanjani & Li, 2014). In sum, when pairs engage in joint revision of their texts, the shared processing of written corrective feedback benefits their W2L as well as W2LL (Storch & Wigglesworth, 2010).

In her five-patterns-model, Storch (2013) identified more and less fruitful patterns of collaboration that tend to emerge when writers work together. These patterns are defined according to two parameters, expressed by two axes, one on equality of contribution, which can be low or high and one on mutuality, which also can be low or high, according to how open the partners are to each other's suggestions (cf. Fig. 1 in Section 4.1). According to Storch (2013) and Watanabe and Swain (2007), the writing processes of *collaborative* pairs in which both partners take shared responsibility and ownership (*high equality* and *high mutuality*), contributing to text generation through deliberation and discussion, are most supportive of L2 learning. To a lesser extent, also *high equality* and *low mutuality* pairs who engage in *cooperative* behaviour (where both partners contribute equally to the final product, without engaging with each other's contributions) and *dominant-dominant* pairs (both stressing their own perspective, leading to potential conflict), as well as *expert-novice* pairs (characterised by *low equality* and *high mutuality)* can establish successful writing processes supporting L2 development. The least advantageous pattern called *dominant-passive* (pairs showcasing *low equality* and *low mutuality)*, is characterized by a lack of deliberation and discussion, which means that this pattern does not benefit L2 learning.

Empirical research into CW has shown that amongst others proficiency plays a major role in establishing positive collaborative patterns. Writers who are more advanced both as writers and L2 learners typically engage in more collaborative patterns that induce languaging and a higher number of LREs (Kim & McDonough, 2008; Storch & Aldosari, 2013). Advanced high-proficiency writers also tend to discuss higher-order concerns during collaborative revision more than low-proficiency writers (Van Steendam et al., 2014).

Research findings also indicate that the type of task influences CW processes. McDonough et al. (2016) found that problem-solving tasks elicit higher numbers of LREs (in comparison to summary tasks), since students engage in higher-order thinking processes and attend to global text problems, such as content, coherence, and text organisation.

While CW was initially investigated in a face-to-face mode, technological innovations of the past decades with affordances supporting shared writing processes (e.g., Google Docs) have sparked new practices of (shared) text generation and an associated line of research looking at (collaborative) writing in an online context. In the next section, we review some of that work that is relevant to our study.

2.2. Web 2.0 and GenAI impacting (collaborative) writing and revision processes

The introduction of web 2.0 tools (e.g., Google Docs, wikis) has made it much easier to work on shared texts, given their affordance for interactive multi-way communication (Luppicini, 2007), thereby transforming writing practices (Hyland, 2016; Vandergriff, 2016). Almost two decades later, digitally mediated CW has become the norm, since it allows pairs or groups to work either synchronously or asynchronously on a shared online text. They can write, elaborate, revise and comment on each other's formulations and content, and then incorporate or reject suggestions they have received from their writing partners in reiterative rounds of text production. In this way, they jointly shape both content and language of the final product, using the commenting function for back-and-forth written deliberations (Zhang et al., 2021).

As Zhang et al. (2021) report in their systematic review, digitally-mediated CW has sparked a wide range of studies investigating diverse aspects of CW from peer interaction and co-composing processes to attitudes towards such tasks. Indeed, in writing pedagogy, the use of digital tools has changed the field, because shared online spaces allow much more process-oriented approaches to teaching and learning (Oh, 2022). This includes the immediate and synchronous provision of online written corrective feedback by a teacher and/or peers during the writing process, which writers or their peers can incorporate instantaneously. For example, Rouhshad and Storch (2016) observed in their study on interplay between the online collaboration mode and collaborative patterns, that synchronous CW leads to truly *collaborative* interaction, contrary to asynchronous CW, which induces more *cooperative* behaviour. Moreover, including the focus on the process not only in teaching and learning, but also in assessment of CW, has a strong positive impact on the intensity of collaboration and the quality of the final products (Zhang & Chen, 2022).

In the past decade, automated feedback on writing provided through AI-based online tools has proliferated and reshaped collaborative writing and revision processes. The recent emergence of GenAI tools has opened up new possibilities for feedback provision to stimulate collaborative engagement in the revision process (Wiboolyasarin et al., 2024). GenAI tools not only can take up the "teacher role" by delivering corrective feedback, but can also act as a "learning peer" (Hwang & Chen, 2023), for instance being prompted with the same writing task as the students to produce a model text for critical comparison with the students' text. This ties in with Woo et al.́s (2024) affirmation that "ChatGPT enables students to write with a machine-in-the-loop", referring to the "collaborative process between a student and a chatbot to complete a writing task" (Woo et al., 2024, Section 1).

In short, GenAI takes digitally-mediated (collaborative) writing and revision to a next level, in that GenAI tools are not merely a 'tool', but can take on a variety of roles across different writing and learning contexts (Gayed et al., 2022). As summarized in Table 1, different scholars have provided classifications and frameworks to pinpoint how GenAI may help (writing) education. Hwang and Chen (2023, p. 2) differentiate between six different roles of GenAI in (writing) education, from functioning as a teacher or tutor to being used as a tool. Warschauer et al. (2023) propose a five-part pedagogical framework to support L2 writers in the process of learning to successfully integrate GenAI into their writing practices. Finally, Su et al. (2023) identify different stages of the writing process (from preparation to reflection) where GenAI can be meaningful. From the short and incomplete overview presented in Table 1 it becomes clear that the manyfold needs for learning opportunities and the new roles for GenAI in education require a reframing and rethinking of tasks for L2 writing.

In the present project, GenAI was integrated as a learning peer or partner (cf. Hwang & Chen, 2023). This is the role that was highlighted for GenAI as potentially conducive to language learning in the L2 classroom by Guo et al. (2022), referring to the capability of GenAI-based chatbots for natural language tasks, such as summarising texts. The task in our project relates to the fourth and fifth part of Warschauer et al.ś (2023) framework and accordingly targets Su et al.ś (2023) reflection stage. More specifically, we invited students to critically assess GenAI-output we had prepared for them in terms of content, structure, and language. We also asked them to reflect and decide what to integrate into their own writing that was produced before accessing GenAI output. By adopting GenAI as a writing buddy in the advanced L2 writing classroom, we aimed to promote awareness of its strengths and weaknesses to achieve students' "calibrated trust" (Ranalli, 2021, p. 3) and enable them to effectively use its output (Kasneci et al., 2023). For example, while GenAI output is known for its linguistic sophistication and grammatical accuracy, users have voiced concerns about the quality of content. In this light, it is important to research the influence of AI generated texts on aspects of L2 writing that go beyond grammatical correctness and target Functional Adequacy (Kuiken & Vedder, 2022).

In addition, having GenAI as a writing buddy by providing writers with two GenAI generated examples of how to synthesise two texts, we stimulated processes of 'inner feedback' (Nicol et al., 2021) during the revision phase of a synthesis writing task. As Yang and Zhang (2010) and Luquin and García Mayo (2024) demonstrate in their studies, use of model texts during revision can facilitate revision processes and have a positive impact on the quality of L2 written production. Yet, it still needs to be explored, whether GenAI can provide teachers and learners with suitable model texts for different contexts and task types, how such model texts might influence dialogues and revision processes, and how they can be used effectively in L2 writing pedagogy.

In our earlier work, Strobl et al. (2024) investigated the revision process students engaged in after a guided comparison of their own texts with AI-generated output in individual and collaborative revision sessions. The guided comparison was effective in facilitating students' critical appraisal of their own texts as well as GenAI texts. The analyses of the revision processes revealed that GenAI output was the main trigger for changes in content whereas Google Docs' automated suggestions triggered the majority of language related adjustments, a finding that ties in with Oh's (2022) observations. Comparing individual and collaborative revision, Strobl et al. (2024) found that the latter context led to a higher percentage of successful revisions. In addition, peer discussions were the most frequent starting point for reviewing content in collaborative settings. These findings triggered the present study, where we zoom in on the collaborative patterns adopted by the pairs during the revision phase to receive a more in-depth perspective.

2.3. *Research questions of the present study*

In the present study, we build on findings from earlier studies on LREs in CW and revision, and on L2 writing in the context of Web2.0 and GenAI technologies. The novelty of our approach lies in the use of the GenAI models and a guided comparison task as triggers to stimulate peer negotiation and LREs, while we refrained from teacher feedback (cf. e.g., Hanjani & Li, 2014). In addition, our students engaged in collaborative revision of a collaboratively (instead of individually) produced text, which means that there is true 'collaborative ownership' of the final product. Within this novel context, we investigated the collaborative patterns in which students engaged in during revision pursuing the following research questions:

Table 1
Different roles of and support by AI for L2 writing. Aspects implemented in our study underlined.

Roles of GenAI in writing Hwang and Chen (2023)	• **teacher/tutor:** providing supplementary materials or examples
	• **student/tutee:** learning from input provided in order to do a task
	• **learning peer/partner:** working as a team member in collaborative learning tasks
	• **domain expert:** providing advice of answers to particular questions
	• **administrator:** summarising findings for decision making
	• **learning tool:** e.g., helping to collect and analyse data
Pedagogical framework to support L2 writers with AI Warschauer et al. (2023)	• **understand:** what the tools can and cannot do
	• **access:** the tools and explore them across different types of tasks
	• **prompt:** the tools appropriately to get the desired output
	• **corroborate:** the output critically in terms of factual correctness and linguistic appropriateness
	• **incorporate:** the output, adapting it to their own needs and writing situation
AI support at different stages of writing Su et al. (2023)	• **preparation:** facilitating idea generation and providing feedback on outlines
	• **editing:** providing feedback on the draft and supplying different perspectives
	• **proofreading:** providing error corrections
	• **reflection:** facilitating reflection through the chat history

RQ1. What **patterns of collaboration** do paired students exhibit when collaboratively revising a text based on a guided comparison of their own jointly written text with models generated by AI?

RQ2. How does joint revision after guided comparison with a GenAI model influence **text quality in terms of Functional Adequacy**?

RQ3. How do students' **patterns of collaboration relate to text quality** in terms of Functional Adequacy?

3. Method

3.1. Context and participants

This study is part of a larger research project on the use of GenAI in L2 German writing and revision activities that consists of several parts, first with focus on individual and then on collaborative writing and revision with GenAI.

After ethical approval and following informed consent, N = 22 students from various BA programs at a university in the Netherlands minoring in L2 German participated in the larger study. They were all enrolled in classes on German language and culture targeting intermediate to advanced proficiency levels (CEFR B2-C1). Data collection sessions took place during the participants' regular seminar hours and the topics and activities aligned with the contents of their course (e.g. popular scientific text forms, German linguistic phenomena, summarising information from various sources into a coherent synthesis). For the present in-depth analysis, we focus on four collaborative pairs (N = 8) who we selected based on practical criteria (quality of the voice recording, full data set of all partners available), with the characteristics presented in Table 2.

Due to availability of computer classrooms, students from different courses worked together during the collaborative writing and revision sessions. Most of the students worked together with a classmate, while some students worked with a partner from a different group with a different proficiency level. Accordingly, students from pairs A, B and D attended the same class, while the two students forming pair C belonged to two different classes (hence, the difference in proficiency). Students within the pairs A, B and D had known each other for at least one year at the time of the intervention. Due to a highly collaborative task-based curriculum adopted in L2 German classes at their institution, all students were acquainted with collaborative writing and joint revision tasks.

3.2. Research design and instruments

In a larger two-week classroom-based intervention, the participants were tasked with writing a synthesis of two popular scientific articles, and to subsequently revise their products based on a guided comparison with GenAI-generated model syntheses (GAIM) of the same source texts. Students worked once independently and once in collaboration with a peer, participating in four writing sessions, that is, the individual writing and revision and the collaborative writing and revision, respectively. In this study we focus on the collaborative revision part (session 4) of this intervention (cf. details below).

The input for the synthesis consisted of two popular-scientific articles on the topic relevant for the study program: anglicisms in German. To create GAIMs for this intervention we used the following prompt in ChatGPT (1) "Write me a synthesis in German of 300–350 words on these two texts. Please make sure that the synthesis contains an introduction, a main body and a conclusion with a final evaluation. The text should be written in an academic style." Prompt (2) was identical but elaborated with "Use *subjunctive I* for indirect quotations." Despite using almost identical prompts both times, texts created by ChatGPT were quite different in content, structure and language use.

As ChatGPT-produced syntheses varied in quality and were not flawless, they served as 'coping models' in our experiment. Zimmerman and Kitsantas (2002) found that coping (peer) models, who made errors in the writing process, had a higher impact on the learning-to-write process of novice writers than mastery models who performed flawlessly. Raedts et al. (2009) confirmed the positive effect of the observation of coping models on the critical thinking abilities of novice syntheses writers. In our intervention, the GenAI *product* (instead of the writing process, like in Zimmerman & Kitsantas, 2002 study) was targeted, and the 'coping model texts' produced by ChatGPT were used in the revision phase instead of the pre-writing phase. Students had to evaluate the coping model texts themselves instead of observing the model in the correction process, like in Zimmerman and Kitsantas (2002). Still, the texts used for comparison in our view can be regarded as 'models' in this sense, because they stimulate critical thinking about fulfilment of task requirements and functional adequacy.

Table 2
Specifications of eight focus participants.

Participant	L1	Gender	Target level of German course	Study background
A1	Polish	f	B2/C1	Art History
A2	Dutch	m	B2/C1	Law
B1	Spanish	f	B2/C1	European Languages and Cultures
B2	Dutch	f	B2/C1	European Languages and Cultures
C1	Dutch	m	B2	International Relations and Organisations
C2	Dutch/Spanish	f	C1	International Relations and Organisations
D1	Dutch	f	C1	European Languages and Cultures
D2	Dutch	f	C1	International Relations and Organisations

To familiarise the students with the target text type and the necessary steps when summarising information from several sources into one coherent text, a 10-minutes explanation video providing explicit instruction about synthesis writing was shown prior to the data collection. The video was based on TRAMPOLINE-strategies for synthesis writing as explained in Buyuktas Kara et al. (2018).

The larger study consisted of four classroom sessions of 90 minutes each. In session 1, participants received two shortened popular scientific articles on *Kiezdeutsch*, a variety spoken by multicultural urban youth communities in contemporary German, and were asked to individually write a synthesis (300–350 words) of the two texts on a computer. During the subsequent session, the students received two GAIMs of the same source articles with the assignment to critically compare their own texts produced in the prior session with the GAIMs and to revise their own texts, integrating the GAIM output where deemed appropriate.

In sessions 3 and 4, this procedure was repeated with two new source articles on anglicisms in the German language and GAIMs of these articles. This time, students were paired up during planning, writing and revision to work collaboratively on a shared digital text and revise them in a joint effort after critically assessing and comparing them to the provided GAIMs.

The process of collaborative revision was guided by an evaluation form with 11 statements concerning the quality of the text (Table 3), drawing on criteria for synthesis writing (Solé et al., 2013; Zhang, 2013). Students had to read the GAIMs, to discuss their opinions and to agree on rating of each text (their own and both GAIMs) with regard to each statement on a 5-point evaluation scale.

The writing and revision assignments were conducted in Google Docs, an environment that the students were familiar with since it is regularly used for writing and peer feedback assignments in their German courses. In addition, participants had full access to any online auxiliaries (e.g. spell check, online dictionaries) while writing and editing their texts. Screen recordings using the software *Screenpresso* captured the students' writing and revision during all four sessions. In session 3 and 4 students audio recorded their deliberation processes with their own smart-phones and submitted these recordings to their class teacher as part of the project.

3.3. *Analysis*

For the subsequent analysis, we worked with verbatim protocols of the discussions of the four focus pairs. Using the software *happyscribe.ai,* we created a draft transcript that was manually corrected and curated by a student assistant.

3.3.1. *Quality and focus of collaborative revision process*

To establish the collaborative patterns of pairs during the writing and revision stages, we followed Storch's (2002, 2013) five-patterns-model for dyadic CW processes with mutuality (i.e., the level of engagement with each other's contributionsa), and equality (i.e., the level of contribution to the task by the two partners; see Fig. 1 in Section 4.1). Two of the authors read the verbatim protocols and independently attributed one of Storch's (2013) categories to the collaborative dialogues, comparing her example dialogues with the dialogues found in the verbatim protocols. Discrepancies between the two raters in two corner cases (i.e., selection of a predominant pattern for dialogues with varying patterns) were resolved through discussion.

In addition to attributing a generic label to each pair, we also identified and counted language-related episodes (LREs) in the verbatim protocols. Swain and Lapkin (1998) define LREs as instances where learners ponder the meaning of linguistic items, choice of grammatical forms, spelling and pronunciation in the process of language output revision. While scrutinising the pair dialogues for LREs, we found a high amount of specifically task-related talk that was not language-related: pairs discussing content selection, text structure, coherence and cohesion and potential issues with so-called 'patchwriting' (Pecorari, 2003), such as excessive textual borrowing bordering on plagiarism. 'Patchwriting' talk both referred to their own actions of copy-pasting GAIM text into their own syntheses and to their observation that the GAIMs contained textual borrowing from the original source texts. Since content selection, text structure, coherence and textual borrowing also are important quality indicators for the completion of the complex task of synthesis writing, we decided to take these episodes into account in our analysis. In coherence with LRE, we coined these episodes 'task related episodes' (TREs). Although LREs are also task-related, and, as such, belong to TREs, we counted them separately to highlight the importance of analysing TREs in addition to LREs when gauging the effect of collaborative patterns on revision in a complex writing task.

3.3.2. *Assessment of students' texts for functional adequacy*

In addition to the process analysis, we also evaluated the products in terms of text quality. Two of the authors rated both the collaboratively written drafts (product session 3) as well as the revised syntheses (product session 4) for dimensions of functional adequacy (FA), that is *content*, *task requirements*, *comprehensibility* and *coherence and cohesion.* FA is defined as "successful task completion by the speaker/writer in conveying a message to the listener/reader" (Kuiken & Vedder, 2022, p. 1). Given the task-based approach of the German language proficiency courses our study was implemented in, the FA scale with its subdivision into four task-related dimensions, made it an appropriate tool for the evaluation of the text quality (in contrast to more accuracy oriented and/or form-based evaluation). Accordingly, using Kuiken and Vedder's (2017) assessment rubric we judged FA on a scale from 1 (lowest) to 6 (highest) for *content* (i.e., the acceptability and consistency of the ideas presented in the students' syntheses), *task requirements* (i.e., the degree to which task purposes and expectations were met), *comprehensibility* (i.e., how much effort a reader might need to understand the writers' intentions), and *coherence and cohesion* (i.e., text flow in terms of argumentation as well as the use of connectors).

In several rounds, both raters double-coded about half of the texts that were presented in a randomised order to mask their attribution to session 3 or 4. Subsequently, ratings were compared and we recalibrated our assessment criteria and ratings by discussing any disagreements. The remaining texts were split into two halves and rated independently. Discrepancies, doubts or questions were resolved through discussion.

Table 3
Tool for the guided collaborative comparison task: Evaluation form for synthesis.

	+ +	+	o	-	–
	1	2	3	4	5
The synthesis reflects the content of the two original texts well.					
The synthesis has a clear and logical structure.					
The introduction summarises the topic of the synthesis.					
Tha main section is divided into clear thematic sections.					
The conclusion clearly rounds off the synthesis.					
The individual topics/thoughts are well interlinked.					
The synthesis reads smoothly in one go.					
The synthesis is reader-oriented: it explains what the reader does not yet know.					
The general use of language is correct.					
The general language usage is varied.					
The language style suits a scientific synthesis.					

1 = completely agree; 2 = agree; 3 = neutral; 4 = disagree; 5 = do not agree at all

4. Results

4.1. Patterns and focus of collaboration during the revision process

To a certain extent, all four focus pairs were found to exhibit collaborative patterns during the revision of their collaboratively produced texts, albeit with differing degrees. While pairs B and C consistently showed collaborative behaviour throughout the whole revision session, in pairs A and D, also episodes of dominant/passive and expert/novice behaviour were observed, respectively. Fig. 1 depicts the distribution of the four pairs in Storch's (2013) model, as attributed by the two researchers.

The numbers of LREs and TREs identified in the verbatim protocols is shown in Table 3. With 18 and 20 LREs respectively, the two pairs with collaborative interaction patterns (B and C) clearly exhibit a higher number of LREs than A (5 LRE) and D (11 LRE). With regards to TREs, the discrepancy between the two groups containing at least one high-proficient participant of the C1 course, C (16 TRE) and D (18 TRE), and the two other groups (A: 4 TRE; B: 3 TRE) is even more considerable.

4.2. Joint revision success in terms of functional adequacy of the product (text quality)

Figs. 2 and 3 present the text quality ratings in terms of functional adequacy (FA, Kuiken & Vedder, 2017) of the collaboratively written (S3) and jointly revised (S4) texts (i.e., revision in S4 based on the guided collaborative comparison with the GAIMs). Fig. 2 shows the development per pair on the different subdimensions of FA, while Fig. 3 depicts calculated mean scores across all four dimensions. The thick line represents the trend, indicating that the collaborative revision based on a comparison with the GAIMs had a positive impact on content choice (Fig. 2 – top left), comprehensibility (Fig. 2 – bottom left), coherence and cohesion (Fig. 2 – bottom right) as well as the overall text quality (Fig. 3). Only fulfilling task requirements (Fig. 2 – top right) suggests an overall decrease following the interaction with the GAIMs.

When focusing on the different pairs, the two collaborative pattern pairings (B and C, represented by the dashed lines in Figs. 2 and 3, respectively) made the larger progress. Group D (expert-novice pattern, high-proficiency) already achieved a very high score of 4.5 out of 6 in S3, but even managed to slightly improve their score in S4 (4.75). The only pair whose synthesis did not improve after revision, was the dominant/passive pair A, whose text in S4 was rated lower (3.5) than the text in S3 (4.25).

Table 5 represents the results of the product analysis separately for each subdimension of FA. Clear and slight improvement between S3 (before revision) and S4 (after revision) is colour-coded in dark and light green shades, respectively, while red shades represent deterioration. This overview suggests that the four pairs attended to different FA subdimensions to different degrees or with differing success during the revision process. In the next section, we will zoom in on the four groups separately to elucidate these differences in revision foci and success by combining process and product related information.

Table 4
Number of language-related episodes (LRE) and task-related episodes (TRE) per collaborative pair.

Pair	LRE	TRE
A	5	4
B	18	3
C	20	16
D	11	18

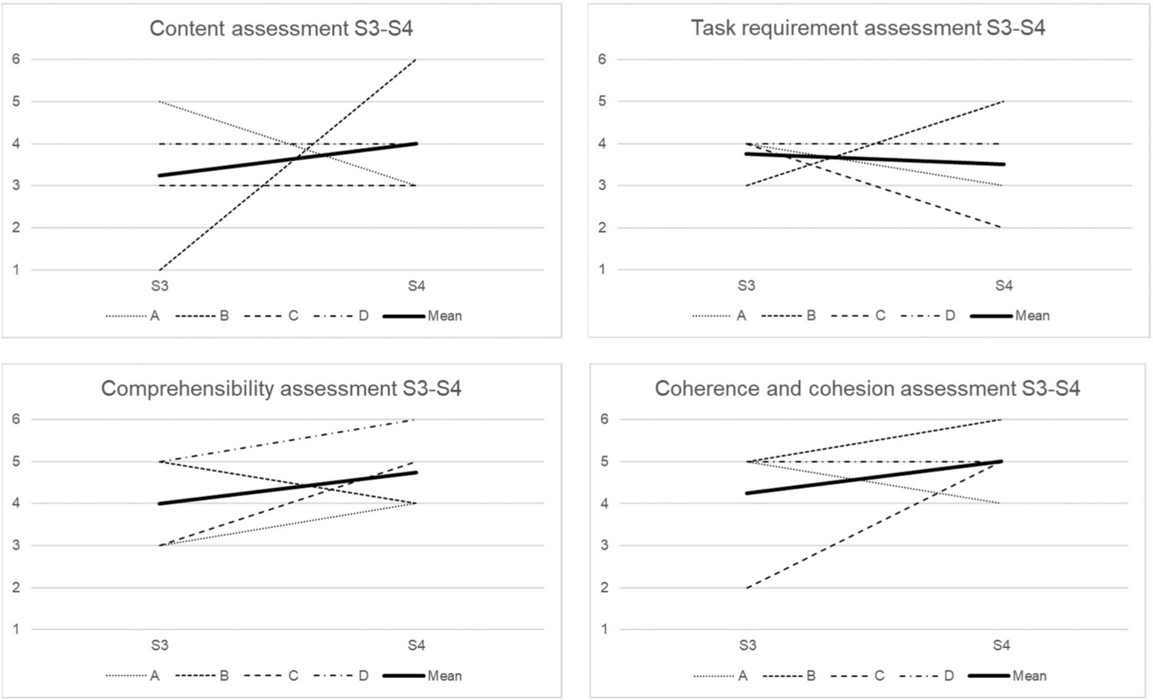

Fig. 2. *Assessment of different dimensions of Functional Adequacy (1 = low to 6 = high) for the collaborative texts before (S3) and after (S4) revision for each of the pairs.*

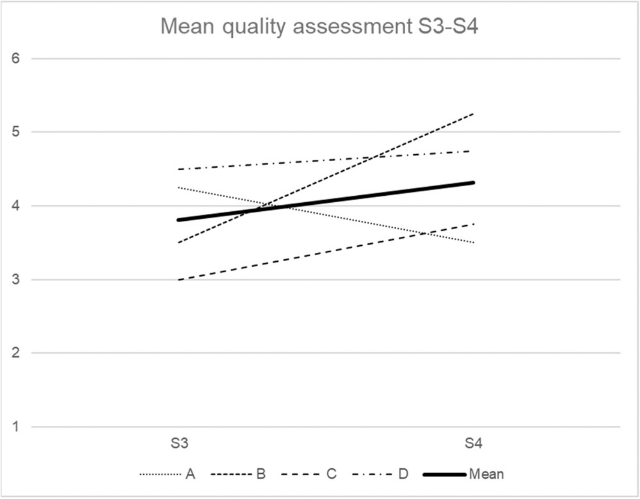

Fig. 3. *Mean scores of Functional Adequacy (1 = low to 6 = high) for the collaborative texts before (S3) and after (S4) revision for all pairs.*

4.3. Relationships between patterns of collaboration (process) and text quality in terms of functional adequacy (product) for each pair

In order to address research question 3, we combine the results of the product measures of the four dimensions with process-related information for each focus pair, providing additional information about their revision focus and quotes from the verbatim dialogues to elucidate the quantitative results.

Pair A showed instances of dominant/passive collaboration patterns. They spent 44 minutes on the revision task, attending to 5 LREs (which is mirrored in their improved score for *comprehensibility* from 3 to 4 out of 6 on the Likert scale) and 4 TREs (three of which were content-related; one related to patchwriting issues). In comparison with the other three groups, they showed little interaction and engagement with each other's suggestions. A1 was mostly typing while A2 watched and sometimes dictated suggested changes. Most

Table 5
FA ratings (Likert scale from 1 to 6) for all four subdimensions, showcasing differences between S3 (before joint revision) and S4 (after joint revision).

	Mean		Content		Task requirements		Comprehensibility		Coherence	
	S3	S4	S3	S4	S3	S4	S3	S4	S3	S4
A	4.25	3.5	5	3	4	3	3	4	5	4
B	3.5	5.25	1	6	3	5	5	4	5	6
C	3	3.75	3	3	4	2	3	5	2	5
D	4.5	4.75	4	4	4	4	5	6	5	5

revision actions consisted in adding fragments from the GAIM in patchwriting style. The following episode found in the verbatim protocol of their collaborative revision dialogue characterises their stance regarding patchwriting.[5]

> A2: Also, *die ChatGPT kopiert, glaube ich, ganze Sätze von den Originaltexten.
>
> _[So, ChatGPT copy-pastes whole sentences of the original texts, I think.]_
>
> _(…)_
>
> _A2: Also, wenn ChatGPT das kopieren kann, können wir das auch. [Well, if ChatGPT can copy-paste, so can we.]_
>
> _A1: Können wir das machen? [Can we do that?]_
>
> _A2: Ja natürlich. [Yes, of course.]_

Pair B maintained a true collaborative interaction pattern throughout the revision process, also taking turns for typing. Their time on task amounted to 76 minutes, which shows their high engagement. On the other hand, the verbatim protocols also revealed a high amount of off-task social talk. They engaged in 18 LREs, but only in 3 TREs (two related to patchwriting; one related to structure). This is surprising, given that their text after revision improved remarkably on the _Content_ (from 1 to 6) and the _Task requirements_ (from 3 to 5) subdimensions. Their dialogue showcases a strong interest in grammatical details and they even talked about learning strategies. This is reflected in the following episode, in which they discuss intricacies of the German grammatical gender.*

> _B1: Ich glaube, es ist_ ‘*DER _Internet’. Stimmt das? [I think it’s’ *DER Internet’. Isn’t it?]_
>
> _B2: Ja, wahrscheinlich, ja. Ja, manchmal, wenn ich ein englisches Wort sehe, dann schreibe ich immer DAS, aber es stimmt nicht. [Oh yes, probably. Sometimes when I see an English word, I always use DAS, but it’s not always correct.]_
>
> _(…)_
>
> _B2: Meiner Meinung nach könnte es auch ‘*DER Internet’ sein. [I think it could also be ‘*DER Internet’.] (looks it up in an online dictionary)_
>
> _B1: Aber ‘DAS Internet’. Es ist auf Niederländisch auch ‘het internet’, habe ich gedacht. Ein ähnliches Wort wie ‘Internet’ war dann plötzlich männlich und das hat mich nicht erstaunt. (…) Aber ‘Internet’, also DAS, okay. Jetzt werde ich das nicht mehr vergessen. [But ‘DAS Internet’. I thought that in Dutch, it‘s also ’het internet. A similar word like ‘internet’ was male [in German] all of a sudden, which I wouldn‘t have expected. (…) But internet, ‘das’, okay. Now, I‘ll never forget it.]_
>
> _(…)_
>
> _B1: Aber so gibt es mehrere dieser Wörter. (…) Also auch ’Moment’ ist schwierig. [But there are more of these words. ‘Moment’ is also difficult.]_
>
> _B2: DER Moment?_
>
> _B1: Ja, ich glaube, wir haben schon eine Diskussion darüber gehabt, ob es DER oder DAS Moment war. Und ich glaube, das ist dasselbe mit ‘Teil’. (…) Aber ich habe keine Ahnung, wann wohl, wann nicht. [Yes, I think we already discussed in class whether it’s DER or DAS Moment. And I think, it’s the same with ‘Teil’. (…) But I have no idea when it’s [male] and when it’s [neutral].]_

* accentuated words (articles) are written in all-uppercase

Pair C was another pair with truly collaborative interaction patterns throughout the revision process. Their lively exchanges were characterised by episodes of joint formulations. With a time on task of 44 minutes, they had a very high amount of LREs (20), which included discussions of minor local concerns, like commas and the German _Umlaut,_ and also a comparably high amount of TREs (16), in which they mainly discussed content issues (10), followed by structure (5) and cohesion (1). Their text mainly improved in terms of _Coherence_ (from 2 to 5) and _Comprehensibility_ (from 3 to 5) after revision. In this pair, we also see a decrease in terms of _Task_

[5] All quotes from the verbatim protocols were translated from original German by the authors (between squared brackets). Non-targetlike language use is marked by an *.

requirements (from 4 to 2). It might be that the GAIM have set them on the wrong foot as the revised version did not reflect the content of the two source texts to an equal extent, which other pairs had criticized about the GAIM. Another possible explanation would be a shift in the focus from *task requirements* to *comprehensibility* and *coherence*. The combined high amount of LREs and TREs results from the fact that, while discussing content and structure of the GAIM and jointly formulating, languaging took place coincidentally, as it were.

C1: 'In diesem Text schauen wir auf…' ['In this text, we look at…']

C2: Vielleicht ohne 'wir', weil das nicht so wissenschaftlich klingt. 'In diesem Text…'

[Perhaps better not to use 'we' since that doesn't sound academic. 'In this text…']

C1: 'In diesem Text schaut auf…' ['This text looks at…']

C2: Oder 'handelt von…' [Or 'is about…']

C1: Ich würde sagen, 'schaut', weil wir probieren zu untersuchen, was der Ursprung dieser Worte ist, zu untersuchen und zu erklären. [I would say 'looks at', because we try to explain the origin of these words.]

C2: Aber wie wissenschaftlich klingt das? Gibt es hier vielleicht noch eine Alternative? 'Diesen Text bezieht sich auf die Frage…' [But is this academic style? Is there a better alternative, perhaps? 'This text deals with the question if…']

Pair D achieved the highest score of all pairs for the collaboratively produced synthesis in S3, that is, before revision (4.5), and still managed to improve after revision (4.75), which aligns with their high proficiency (both C1). Interestingly, while both are high-achieving C1 students, they still adopted an expert/novice pattern in some instances of the revision process (46 minutes). Improvement was achieved on the *Comprehensibility* dimension (from 5 to 6), which mirrors their occasional LREs (11) including discussions about lexical choices. This pair took 48 minutes to complete the revision and excelled mainly in terms of their high amount of TREs: nine related to content, five to structure, three to cohesion and one to patchwriting. Even if they expressed satisfaction with their text produced in S3, they still scrutinised the GAIMs, trying to integrate fragments into their revised text. The following episode showcases their preoccupation with academic writing style, as well as the expert/novice pattern observed.

*D1: Was mir schon wieder aufgefallen war, ist, ChatGPT hat wieder *kein Konjunktiv 1 benutzt. Gar keinen. [I noticed again that ChatGPT didn't use the Konjunktiv 1. None at all.]*

D2: Nein? [Really?]

*D1: Nein, gar keinen. Weil Konjunktiv 1 ist zum Beispiel 'man nehme'. Und diese Formen gibt es nicht. (…) Man *sollten *es hier eigentlich sagen: 'Während der erste Text die Notwendigkeit von Sprachwissenschaft betone und argumentiere, dass…', aber das macht *es nicht. [Nope. Because Konjunktiv 1 is for example 'man nehme'. And there are no such constructions here. So, here it should be: 'Während der erste Text die Notwendigkeit von Sprachwissenschaft betone und argumentiere, dass…', but it's not.] D2: Und das *hat es wirklich schöner gemacht. [And that would have made it a lot nicer, indeed.]*

*D1: Das *hat es auch wissenschaftlicher gemacht. [And it would have also made it more academic.]*

5. Discussion

This study set out to investigate collaborative revision patterns in pairs of university students working on a jointly written synthesis text in German L2. As the main intervention, we provided students, after their initial collaborative writing of a synthesis text, with two model syntheses created by GenAI (GAIM). Through guided comparison of their own text with the GAIMs, we intended to trigger processes of inner feedback (Nicol et al., 2021), deliberative dialogue, foster their awareness of their own L2 writing capabilities, and highlight the benefits and pitfalls of using GenAI to support their L2 writing. In an earlier analysis, Strobl et al. (2024) looked at the patterns of revision behaviour in the larger group of participants (N = 22). In this study, we focus on four pairs of students (N = 8) by (a) analysing their discussions during the revision process and (b) assessing the quality of their texts before and after the guided comparison in terms of functional adequacy (FA, Kuiken & Vedder, 2017).

5.1. Patterns of collaborative revision behaviour

Our first research question asked what patterns of collaboration paired students exhibited when collaboratively revising a text based on a guided comparison of their own jointly written text with models generated by AI. Of our four pairs, two (B and C) adopted a clear collaborative pattern of interaction during revision with high mutuality and equality. Their interactions were characterised by engaged discussions on both language- and in case of pair C also task-related aspects of their assignment, that is, high numbers of language and task related episodes (LREs and TREs, respectively). The fact that groups B and C exhibit the highest amount of LREs during the collaborative revision session corresponds with earlier findings that pairs with truly collaborative interaction patterns are more likely to produce LREs than pairs with different interaction practices during writing (Storch, 2002; Storch & Aldosari, 2013). It seems that in our sample, this tendency is further augmented because of the participants' language proficiency level, which was high and allowed them to engage in extensive languaging even in their L2 (Kim & McDonough, 2008).

A slightly different case was Pair D, which showed patterns of expert/novice collaboration and did not generate a lot of LREs. Probably, in this pair fewer LREs occurred because they were already performing at a very high linguistic level (C1 proficiency). Accordingly, their discussions mainly pertained to the task, as manifested by a high number of TREs. Indeed, many more TREs were observed in the two pairs with at least one high-proficient C1 participant (C and D). As such, our findings tie in with research on CW and revision that attributed highly proficient L2 writers an attention shift from lower-order aspects of textual quality, like spelling and grammar, typically observed in low-proficient writers, towards higher-order aspects of content, structure and coherence (e.g., Van Steendam et al., 2014). It is noteworthy, that these deliberative dialogues seem to be prominent also during revision processes triggered by comparisons with GAIMs.

Interestingly, earlier research identified task effects in collaborative writing such that problem-solving tasks elicit higher numbers of LREs than summary writings (McDonough et al., 2016). Our guided comparison could be seen as a problem-solving task that elicits deliberative dialogue (cf., Casado-Ledesma et al., 2021) when comparing a previously produced synthesis with GAIMs. Hence, it required our students to evaluate differences in quality of their own and GAIM texts and asked for a decision on whether and how to revise their own products. This type of task triggered deliberation on both language use as well as content and text organisation. Accordingly, we found that our students did not only engage in LREs, but also in TREs. The latter entails attention to content selection, text structure, coherence and cohesion as well as potential problems with excessive textual borrowing (i.e., patchwriting, Pecorari, 2003) - all points that indeed were manifested in the discussions between our pairs (cf. excerpts presented in Section 4.3).

In light of the three pairs working together at high levels of mutuality (B, C, D), it is surprising to see the lack of interaction in pair A. This pair, which showed patterns of a dominant/passive collaboration, hardly engaged in any LREs nor TREs. Their collaboration probably was dominated by an attitude of ‘getting the task done’, with one pair member typing and the other making limited suggestions without engaging in shared reflection or elaborated deliberations. Speculating about the why, one reason might be that these two particular L2 writers were students of *Law* and *Art History*, respectively. Accordingly, they were enrolled in German classes as an elective rather than as a compulsory part of their studies (for the other students, enrolled in *European Languages and Cultures* or *International Relations and Organisations*, studying a language is an integral part of their academic training) and the topics of the intervention tasks were not related to their main field of studies. One might argue that already from the outset, Pair A’s interest in and attitude towards the content- and language-related aspects of writing in German L2 was less inherently linked to their personal repertoire. Probably, they were, therefore, less inclined to engage in languaging practices. It would be interesting to dedicate future research to such individual differences impacting CW and revision processes, taking into account writing space variables related to the curriculum, the writing environment and individual writing motivation and experiences, a suggestion recently made by Fogal (2024) to expand the CW research framework. Based on our current data, we can only support Storch’s (2009) claim that L2 learners’ individual characteristics can influence the degree of collaboration when writers are paired, and, in line with Van Steendam et al. (2014), Kim and McDonough (2008), and Casado-Ledesma et al. (2021), it seems that proficiency level and task type are likely to mediate the focus of a pair’s deliberation processes (LREs vs. TREs).

5.2. Text quality effects of guided comparison with a GenAI model

Our second research question focused on how joint revisions after guided comparison with a GenAI model might influence text quality in terms of Functional Adequacy (FA, Kuiken & Vedder, 2017, 2022). Our data show that a collaborative L2 text revision task with GenAI as a third player has increased overall text quality of the L2 writers’ syntheses, in particular, related to the subdimensions of content, text comprehensibility, and cohesion and coherence.

The analyses of students’ deliberation dialogues during collaborative revision revealed that working with GAIMs as a reference encouraged them to not only scrutinise their own products (cf., ‘inner feedback’, Nicol et al., 2021), but also helped them to take a critical stance towards the GAIM texts while engaging in deliberative dialogues (Casado-Ledesma et al., 2021). The discussions showed substantial critical engagement with both the quality of language and content that the GAIMs provided as well as with ways to use GAIMs for improving their own syntheses. These processes, in turn, seem to have elicited language- and task-related episodes (LREs and TERs) between paired L2 writers, with all the benefits discussed in the former section and earlier work (see Storch’s, 2022 review). In line with Casado-Ledesma et al. (2021) it seems that the explicit instruction (i.e., the guided comparison) helped students to focus on relevant aspects of their own and GAIM texts that helped them to improve their writing through revision.

Interestingly, when looking into FA subdimensions, our data show that only pair B was able to increase their score on task requirements, while other pairs demonstrated no (pair D) or even a decrease in meeting the task demands (pair A and C). Even pairs that improved substantially in their overall text quality lowered their score on this dimension (e.g., C). A possible explanation might be that the short explanatory video prior to the data collection sessions was not sufficient to introduce the target text format (synthesis) and raise awareness of text-specific requirements such as a balanced representation of both source-texts’ contents. It might also be that the choice of GenAI-created models had a negative influence: in order to spark a critical engagement with the GAIMs, we deliberately included AI generated models that did not fulfil all the task-specific demands. Students being insecure about the specific text type, might have been influenced negatively by models that did not meet task requirements, which then led to an overall decrease on this FA subdimension. More detailed analyses of students’ engagement with the GAIMs may offer some insights into this aspect of FA in our data.

For now, it seems that adding GAIMs as a third entity to CW and revision tasks holds the potential to trigger students’ engagement with their own text and foster inner feedback (Nicol et al., 2021), to promote noticing both at a linguistic and a task-related level (Schmidt, 1990; Storch, 2022), and to sensitise L2 writers for the strengths and weaknesses of GenAI-created output (Kasneci et al., 2023; Ranalli, 2021).

5.3. The relationship between collaboration patterns and text quality

The third and final research question of the present study aimed at looking into the relationship between patterns of collaboration and text quality in terms of Functional Adequacy (FA). Our data reveal that the highly mutual collaborative pairs B and C achieved the greatest overall improvement of text quality, while the less efficient dominant/passive pair A did not increase their text quality and even included some revisions that made their synthesis worse.

For the collaborative pairs and in line with earlier work (e.g., Donato, 1994; Hanjani & Li, 2014), the deliberations during LREs showcase effective instances of collaborative scaffolding (e.g., excerpt pair B discussing the correct gender of a German noun) where both writers pool their knowledge together to come to higher-level language use (Storch, 2022). In cases where the pairing included a student with C1 proficiency (pair C and D), the deliberations clearly went beyond lower-order processes (e.g., spelling), but addressed topics that would improve the text in terms of genre (e.g., excerpt pair C, discussing academic language use; excerpt D, discussing a verb tense related to academic language) - as was found by Van Steendam et al. (2014). In our coding, such discussions were more often assigned to TREs than LREs because they aimed at meeting task requirements (e.g., balancing information from two sources in the synthesis) and/or improving coherence. In the assessment of FA, this resulted in higher text quality ratings for task requirements, comprehensibility (pair B) and coherence (pair B and C). In contrast, the pattern of pair A, demonstrating a lack of joint efforts to engage in evaluative or social deliberations as identified by Hanjani and Li (2014) carry the risk of revisions resulting in lowering text quality (e.g., the absence of LREs and TREs in pair A leading to lower FA ratings).

5.4. The value of GenAI and GenAI models

The detailed analysis of pair talk as the writers were discussing their intended revisions also showed several benefits and drawbacks of the online environment. On the one hand, as we see in pair B, the availability of an online dictionary (e.g., to quickly look up the gender of a noun) provides an immediate answer to questions arising during writing and revision activities, which entails that digitally mediated L2 writing and feedback shifts from a product- to a more process-oriented activity (Oh, 2022). On the other hand, Strobl et al. (2024) found in an earlier analysis that L2 writers relied to a large extent on the automatic spelling correction and grammatical feedback by Google Docs. Similarly, collaborating pairs often followed and/or accepted suggestions for the continuation of a phrase or formulations offered by Google Docs without any critical evaluation of the suggested content or wording. In contrast, the guided comparison of students' own texts with GAIMs, that formed the heart of the current intervention, triggered many effective joint revision processes and engaged the L2 writers in fruitful deliberations with increased text quality as a result. It suggests that the explicit instruction on how to make use and evaluate the value of the affordances provided by technology might be crucial for student learning. This aligns with anecdotal evidence of post-task talks that indicated that students valued the intervention as it had increased their awareness of how to evaluate GenAI text and work with it in the future.

In short, adding GenAI into the L2 writing classroom offers many opportunities for beneficial processes of L2W and W2LL (Manchón, 2011) revision processes - if implemented in an effective pedagogic way, which we will discuss below.

5.5. Limitations and implications

This study has shown that integrating GAIMs as references for text revision into L2 teaching represents a promising tool to (1) foster learners' critical engagement with their own texts, (2) raise students' awareness of how to use GenAI output effectively to improve their products, and (3) sensitise learners for GenAI's strengths and shortcomings. Yet, our work also has its limitations, of which we will highlight the following pertaining to theory, methodology and teaching practice. Research into GenAI has taken a flight ever since OpenAI released ChatGPT. Also for L2 writing, the number of theoretical and empirical studies becoming available is growing at an unprecedented rate. Consequently, it is virtually impossible to present an up-to-date theoretical background, and we are aware that by the time this paper goes into press, many new insights will have been published, which will not have been incorporated. Based on our current review, we consider that our research does form a valuable contribution to the field, not least because we are working on German as an L2 and thereby help counter the bias of studies on English writing only (Kubota, 2022).

Methodologically, we have chosen to evaluate text quality with a focus on Functional Adequacy (Kuiken & Vedder, 2022) only, leaving out other means of assessment focusing on accuracy on its own or the CAF-triad of Complexity, Accuracy, and Fluency (Wu & Michel, 2024). Given the task-based approach of the specific classroom under investigation, we deem the FA ratings appropriate, yet employing a wider range of evaluation criteria would undeniably generate other valuable insights.

Finally, as with any classroom-based study, our data collection in this ecologically valid context meant that out of an initial pool of twenty-two students, only four pairs (the ones we discuss in the present study) had participated in all classes over the course of 2 weeks that this study needed for its implementation. A larger data-set would allow for more generalizable implications, yet our fine-grained examination has generated interesting insights that seem to be transferable to other similar settings. An additional open question to be investigated in future research is, whether similar effects can be observed in L2 learners with lower proficiency levels.

By performing the current in-depth analyses of the four focus pairs, we were able to identify patterns of collaboration that provide valuable insights for research and teaching with GenAI in an L2 writing class. In particular, our study has identified some pedagogical insights to take into account when planning interventions with GAIMs, of which we will highlight the following. First, we found that group constellations and individual collaborative characteristics impacted the level of engagement students demonstrated with their own product and the GenAI model texts. For future implementations of similar interventions drawing on collaborative writing and revision, it is therefore advised to carefully consider the pairings of students. Inefficient constellations (in our study of the dominant/

passive type) might diminish intended learning effects. Second, it seems that students benefit from substantial introduction to the task-specific requirements (in our case, synthesis writing) prior to interactions with GAIMs. Ensuring that students know what to expect, will presumably trigger most beneficial and critical comparisons of own text with GAIMs. Third, our focus on task-related episodes (TREs) established itself as a fruitful approach to code interactions elicited by task-based collaborative writing tasks. We found that in addition to LREs, task-related deliberations engaged L2 writers in higher-order thinking processes, raising their awareness of global text concerns like content choice, coherence, and text organisation.

6. Conclusion

When GenAI (e.g., ChatGPT) was introduced to the larger public it created an outcry in education, because educators across the globe feared that students will stop engaging with content and language during writing in ways that support their learning. Indeed, GenAI has transformed L2 writing practices and, as our study shows, also CW and revision has undergone major changes. In this light, our study presents an innovative intervention, where students of German L2 are asked to compare their own collaboratively written text with two AI generated models. Findings revealed that our intervention indeed led to students engaging in fruitful deliberation processes making them aware of both language-related and task-related features they should learn to critically evaluate when working with GenAI to support their L2 writing. As such, this article is an example on how the teaching of L2 writing can embrace GenAI as a new partner in supporting L2 learners to become good writers while also learning a new language.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Abel Niklas: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Bazhutkina Iryna:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Strobl Carola:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Michel Marije:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Funding, Formal analysis, Conceptualization.

Data availability

The authors do not have permission to share data.

References

Buyuktas Kara, M., Van Steendam, E., Rijlaarsdam, G., & Kuru, H. (2018). The effect of two modes of strategy instruction: Modeling vs. Presentational. *International Online Journal of Education and Teaching (IOJET), 5*(2), 460–495. ⟨http://iojet.org/index.php/IOJET/article/view/313/247⟩.

Casado-Ledesma, L., Cuevas, I., Van den Bergh, H., Rijlaarsdam, G., Mateos, M., Granado-Peinado, M., & Martín, E. (2021). Teaching argumentative synthesis writing through deliberative dialogues: instructional practices in secondary education. *Instructional Science: An International Journal of the Learning Sciences, 49*(4), 515–559. https://doi.org/10.1007/s11251-021-09548-3

Cumming, A. (2020). L2 writing and L2 learning. In R. M. Manchón (Ed.), *Writing and Language Learning: Advancing research agendas, 56* p. 29). John Benjamins. https://doi.org/10.1075/lllt.56.02cum.

Donato, R. (1994). Collective scaffolding in second language learning. In J. P. Lantolf, & G. Appel (Eds.), *Vygostkian approaches to second language research* (pp. 33–56). New Jersey: Ablex.

Fogal, G. G. (2024). Expanding the collaborative writing research framework: A longitudinal analysis of how collaborative and independent writers orient to writing spaces. *Journal of Second Language Writing, 63*, Article 101096. https://doi.org/10.1016/j.jslw.2024.101096

Gayed, J. M., Carlon, M. K. J., 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*, Article 100055. https://doi.org/10.1016/j.caeai.2022.100055

Guo, K., Wang, J., & Chu, S. K. W. (2022). Using chatbots to scaffold EFL students' argumentative writing. *Assessing Writing, 54*, Article 100666. https://doi.org/10.1016/j.asw.2022.100666

Hanjani, A. M., & Li, L. (2014). Exploring L2 writers' collaborative revision interactions and their writing performance. *System, 44*, 101–114. https://doi.org/10.1016/j.system.2014.03.004

Hwang, G.-J., & Chen, N.-S. (2023). Editorial position paper: Exploring the potential of generative artificial intelligence in education: Applications, challenges, and future research dir*ections. Educational Technology Society, 26*(2), I–XVIII. https://doi.org/10.30191/ETS.202304_26(2).0014

Hyland, K. (2016). Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing, 31*, 58–69. https://doi.org/10.1016/j.jslw.2016.01.005

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, Y., & McDonough, K. (2008). The effect of interlocutor proficiency on the collaborative dialogue between Korean as a second language learners. *Language Teaching Research, 12*(2), 211–234. https://doi.org/10.1177/1362168807086288

Kormos, J. (2023). The role of cognitive factors in second language writing and writing to learn a second language. *Studies in Second Language Acquisition, 45*(3), 622–646. https://doi.org/10.1017/S0272263122000481

Kubota, R. (2022). Decolonizing second language writing: Possibilities and challenges. *Journal of Second Language Writing, 58*, Article 100946. https://doi.org/10.1016/j.jslw.2022.100946

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

Kuiken, F., & Vedder, I. (2022). The assessment of functional adequacy in language performance. *TASK, 2*(1), 1–7. https://doi.org/10.1075/task.21009.kui

Luppicini, R. (2007). Review of computer mediated communication research for education. *Instructional Science, 35*(2), 141–185. https://doi.org/10.1007/s11251-006-9001-6

Luquin, M., & García Mayo, M. d P. (2024). A longitudinal study of the effects of model texts on EFL children's written production. *System, 120*. https://doi.org/10.1016/j.system.2023.103190

Manchón, R. M. (2011). *Learning-to-Write and Writing-to-Learn in an Additional Language*. John Benjamins Publishing Company. 〈http://digital.casalini.it/9789027284839〉.

McDonough, K., Crawford, W. J., & De Vleeschauwer, J. (2016). 7. Thai EFL learners' interaction during collaborative writing tasks and its relationship to text quality. *Language Learning and Language Teaching*, 185–208. https://doi.org/10.1075/lllt.45.08mcd

Nicol, D., Quinn, N., Kushwah, L., & Helen Mullen, C. M. B. E. (2021). Helping learners activate productive inner feedback: Using resource and dialogic comparisons. *LTSE, 44*.

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*, Article 100916. https://doi.org/10.1016/j.jslw.2022.100916

Pecorari, D. (2003). Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing, 12*(4), 317–345.

Raedts, M., Daems, F., Van Waes, L., & Rijlaarsdam, G. (2009). Observerend leren van peer models bij een complexe schrijftaak. *Tijdschrift voor Taalbeheersing, 31*(2), 142–165.

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*, Article 100816. https://doi.org/10.1016/j.jslw.2021.100816

Rouhshad, A., & Storch, N. (2016). A focus on mode: Patterns of interaction in face-to-face and computer-mediated contexts. In M. Sato, & S. Ballinger (Eds.), *Peer interaction and second language learning: Pedagogical potential and research agenda* (pp. 267–289). John Benjamins. https://doi.org/10.1075/lllt.45.11rou.

Sasaki, M. (2023). AI tools as affordances and contradictions for EFL writers: Emic perspectives and L1 use as a resource. *Journal of Second Language Writing, 62*, Article 101068. https://doi.org/10.1016/j.jslw.2023.101068

Schmidt, R. W. (1990). The role of consciousness in second language learning1. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

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*, 63–90. https://doi.org/10.1177/0741088312466532

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning, 52*(1), 119–158. https://doi.org/10.1111/1467-9922.00179

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing, 18*(2), 103–118. https://doi.org/10.1016/j.jslw.2009.02.003

Storch, N. (2013). *Collaborative Writing in L2 Classrooms*. Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781847699954

Storch, N. (2018). Written corrective feedback from sociocultural theoretical perspectives: A research agenda. *Language Teaching, 51*(2), 262–277. https://doi.org/10.1017/S0261444818000034

Storch, N. (2019). Collaborative writing. *Language Teaching, 52*(1), 40–59. https://doi.org/10.1017/S0261444818000320

Storch, N. (2022). Theoretical perspectives on L2 writing and language learning in collaborative writing and the collaborative processing of written corrective feedback. In R. M. Manchón, & C. Polio (Eds.), *The Routledge Handbook of Second Language Acquisition and Writing* (pp. 22–34). Routledge. https://doi.org/10.4324/9780429199691-10.

Storch, N., & Aldosari, A. (2013). Pairing learners in pair work activity. *Language Teaching Research, 17*(1), 31–48. https://doi.org/10.1177/1362168812457530

Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake, and retention of corrective feedback on writing: Case Studies. *Studies in Second Language Acquisition, 32*(2), 303–334. https://doi.org/10.1017/S0272263109990532

Strobl, C., Menke-Bazhutkina, I., Abel, N., & Michel, M. (2024). Adopting ChatGPT as a writing buddy in the advanced L2 writing class. *Technology in Language Teaching Learning, 6*(1), 1–19. https://doi.org/10.29140/tltl.v6n1.1168

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, Article 100752. https://doi.org/10.1016/j.asw.2023.100752

Swain, M. (2006). Languaging, Agency and Collaboration in Advanced Second Language Proficiency. In H. Byrnes (Ed.), *Advanced Language Learning: The Contribution of Halliday and Vygotsky* (pp. 95–108). Continuum.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass, & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *Modern Language Journal, 82*, 320–337. https://doi.org/10.1111/j.1540-4781.1998.tb01209.x

Vandergriff, I. (2016). *Second-Language Discourse in the Digital World: Linguistic and Social Practices in and beyond the Networked Classroom*. John Benjamins.

Van Steendam, E., Rijlaarsdam, G. C. W., Van den Bergh, H. H., & Sercu, L. (2014). The mediating effect of instruction on pair composition in L2 revision and writing. *Instructional Science, 42*(2014), 905–927. https://doi.org/10.1007/s11251-014-9318-5

Verspoor, M. (2017). Complex dynamic systems theory and L2 pedagogy: Lessons to be learned. In L. Ortega, & Z. Han (Eds.), *Complexity Theory and Language Development: In celebration of Diane Larsen-Freeman* (pp. 143–162). John Benjamins. https://doi.org/10.1075/lllt.48.

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of english as a second or foreign language. *Journal of Second Language Writing, 62*, Article 101071. https://doi.org/10.1016/j.jslw.2023.101071

Watanabe, Y., & Swain, M. (2007). Effects of proficiency differences and patterns of pair interaction on second language learning: Collaborative dialogue between adult ESL learners. *Language Teaching Research, 11*(2), 121–142. https://doi.org/10.1177/136216880607074599

Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12819-4

Wiboolyasarin, W., Wiboolyasarin, Y., Suwanwihok, K., Jinowat, N., & Muenjachoey, R. (2024). Synergizing collaborative writing and AI feedback: An investigation into enhancing L2 writing proficiency in wiki-based environments. *Computers and Education: Artificial Intelligence, 6*. https://doi.org/10.1016/j.caeai.2024.100228

Wu, M., & Michel, M. (2024). Complexity-accuracy-lexis-fluency (CALF) as a pedagogical target. In C. A. Chapelle, & M. Sato (Eds.), *The Encyclopedia of applied linguistics* ((second edition): Instructed second). language acquisition. Wiley-Blackwell.

Yang, L., & Zhang, L. (2010). Exploring the role of reformulations and a model text in EFL students' writing performance. *Language Teaching Research, 14*(4), 464–484. https://doi.org/10.1177/1362168810375369

Zhang, C. (2013). Effect of instruction on ESL students' synthesis writing. *Journal of Second Language Writing, 22*, 51–57. https://doi.org/10.1016/j.jslw.2012.12.001

Zhang, M., & Chen, W. (2022). Assessing collaborative writing in the digital age: An exploratory study. *Journal of Second Language Writing, 57*, Article 100868. https://doi.org/10.1016/j.jslw.2022.100868

Zhang, M., Gibbons, J., & Li, M. (2021). Computer-mediated collaborative writing in L2 classrooms: A systematic review. *Journal of Second Language Writing, 54*, Article 100854. https://doi.org/10.1016/j.jslw.2021.100854

Zimmerman, B., & Kitsantas, A. (2002). Acquiring writing revision and self-regulatory skill through observation and emulation. *Journal of Educational Psychology, 94*, 660–668.

Marije Michel (PhD Applied Linguistics, University of Amsterdam) is professor of *Second Language Acquisition* (SLA) at Groningen University (the Netherlands). Her research and teaching focus on cognitive and social aspects of SLA and Task-based Language Teaching (TBLT). Her work is known for mixed-methods approaches to investigating second language processing. For example, in her recent work she has used eye-tracking and key-stroke logging to investigate second language writing processes and alignment in digitally mediated communication. Marije is member of the executive committee of the *European Second Language Association (EuroSLA),* the *International Association TBLT* and the Netherlands-based *English Academy for Newcomers*.

Niklas Abel is a PhD candidate at the Amsterdam University of Applied Sciences and the University of Amsterdam, as well as a lecturer for German language proficiency at Groningen University (the Netherlands). His research focuses on language teaching in linguistically and culturally diverse contexts, with a specific interest in Dutch pre-vocational education. In his language teaching, he applies communicative, usage-based approaches and has a special interest in implicit language instruction, Task-Based Language Teaching (TBLT), Content and Language Integrated Learning (CLIL), Film and Language Integrated Learning (FLIL) as well as the effect of virtual exchange on the development of language proficiency and intercultural competences.

Iryna Bazhutkina is a PhD candidate at the University of Antwerp and the University of Groningen. In her PhD project she investigates the development of pragmatic competence in L2 German in different contexts and analyses how individual differences influence this development. Before starting her doctoral study, she worked as a lecturer of German at the Department of European Languages and Cultures at the University of Groningen. In her teaching and research, she is interested in Task-Based Language Teaching (TBLT), Content and Language Integrated Learning (CLIL), Film and Language Integrated Learning (FLIL), the role of technology in language education and citizenship education in language teaching.

Carola Strobl is associate professor for *Translation and Applied Linguistics* at the Department of Applied Linguistics, Translation and Interpreting Studies, University of Antwerp, Belgium. She has a teaching career at universities in Germany, Italy, Portugal, and Belgium. Her research is situated in the areas of instructed second language acquisition and translation studies. She conducts corpus-based research of learner language and translated language, with a focus on cohesion-building strategies. Furthermore, she takes a special interest in second language writing and translation pedagogies, exploring the potential of online technologies in the learning process.