A Corpus Linguistic Perspective on Error Detection and Correction in English Academic Texts. There's an AI for That?

Katharina Deckert (University of Bamberg)

submitted 22 December 2025, accepted 30 September 2026, published 29 October 2026
https://doi.org/10.62408/ai-ling.v5i1.68

katharina.deckert(at)uni-bamberg.de

Abstract
Students and researchers from various fields of study and from different language backgrounds use English as a *lingua franca* when writing academic texts. However, even advanced and proficient users continue to make errors (Lennon 1991; Uzuner 2008). Therefore, the arrival of new AI writing tools might seem like the ideal solution, especially for non-native users of English struggling with writing academic texts. The goal of this study is to provide an overview of how useful AI tools can be for error correction purposes from a linguistic perspective. A representative selection of typical German learner errors from different linguistic problem areas was identified using a corpus linguistic approach. Nine AI tools were used to correct fictitious learner texts containing these errors and results indicate notable discrepancies in terms of the tools' *detection accuracy* and *correction suitability*. Critical evaluation of AI applications is one aspect of *AI literacy* (D. T. K. Ng et al. 2021, OECD 2025), a future skill needed to use these technologies responsibly and competently. This paper provides a basis for further research in this rapidly developing field and offers practical insights for users of English who wish to leverage AI tools as *artificial proofreaders* for their academic texts.

Keywords
academic writing, error correction, corpus linguistics, AI tools, AI literacy

1 Introduction

Sharing one's findings with the research community is a necessary step scholars are required to take after having conducted their research. However, the process of writing and publishing a scientific paper can be frustrating as various factors must be considered before, during and after the writing process (Gilmore et al. 2006).

One important and complex choice writers are required to make is the decision of which language to publish their research in (López-Navarro et al. 2015). There are various reasons why non-native scholars might decide to publish in English instead of their L1 (Uzuner 2008; López-Navarro et al. 2015), such as wanting to reach a wider audience or to publish in a specific journal (Flowerdew 1999a; Sheldon 2020). These reasons can "emerge from both personal aspirations

Deckert, Katharina. 2026.
A Corpus Linguistic Perspective on Error Detection and Correction in English Academic Texts. There's an AI for That?
AI-Linguistica. Vol. 5 No. 1
DOI: 10.62408/ai-ling.v5i1.68
ISSN: 2943-0070

and institutional imperatives" (Li & Wang 2026: 22). Especially for findings that are deemed appropriate and important for an international audience, it is perceived as more useful to publish one's work in English (López-Navarro et al. 2015), which is why researchers from various fields of study and from different language backgrounds tend to use English as a *lingua franca* (Ammon 1996; Curry & Lillis 2007; Flowerdew 2019; Amano et al. 2023) even though "the majority of the world's scholars do not possess English as their first language" (Flowerdew 2008: 77).

This can be challenging, especially for non-native researchers trying to keep up with international competitors (Amano et al. 2023). Therefore, innovative technologies from the field of *Artificial Intelligence*, including both generative AI applications as well as dedicated AI writing tools, might seem like perfect *artificial proofreaders*. But how useful are AI tools for text correction? The goal of the present study is to provide an overview of how effective AI tools can be for this purpose.

2 Literature review

2.1 English as an academic *lingua franca*

As Flowerdew (1999a: 124) observes, "English has increasingly come to be accepted as the dominant language for publication of academic research findings". This "means that scholars around the world are under increasing pressure to publish their research in English" (Curry & Lillis 2007: 6), which poses a variety of challenges, especially for non-native users of the language. For example, writing in another language requires more time and is perceived as more difficult (Flowerdew 1999b; Ramírez-Castañeda 2020), which is why feeling pressured to publish in a language that is not one's L1 can cause difficulties, dissatisfaction and anxiety (Hanauer & Englander 2011; Hanauer et al. 2019).

Even advanced and proficient users continue to make errors when writing in English, be it as a second language (ESL), additional language (EAL) or foreign language (EFL) (Lennon 1991; Uzuner 2008; Flowerdew 2019). While these deviations from the norm might not necessarily be considered as entirely erroneous but rather as features that native speakers would not use in a similar context (Lennon 1991), when competing in an international context to publish one's article in a high-ranking journal even small inaccuracies can influence the decision of which article will be published and which will not (Hyland 2016b; McKinley & Rose 2018; Strauss 2019; Politzer-Ahles et al. 2020). Although non-standard "lingua franca English is increasingly accepted, there will still be certain standards of intelligibility and clarity that need to be met. It is not a case of 'anything goes'" (Flowerdew 2019: 258), so sharing one's research in a foreign language "needs intensive language studies and remains, as a rule, time-consuming and error-stricken" (Ammon 1996: 261). Therefore, non-native researchers must find a way to

overcome these language barriers (Amano et al. 2023) and have their language errors corrected, for example "by native speakers who have to be paid for it" (Ammon 1996: 260).

While this linguistic injustice (summarised in Politzer-Ahles et al. 2020) has been called into question because writing in the academic register is not acquired automatically but needs training in native speakers as well (Hyland 2016a, 2016b), the present paper takes a similar position to Flowerdew's (2019), who points out that "there is a certain set of challenges in writing for publication which are shared by EAL [English as an Additional Language] and L1 writers, but EAL writers have an additional set of linguistic challenges, which do not apply (to such an extent) to L1 writers" (257). Therefore, the focus of this paper is on non-native users of English.

2.2 Error detection and correction

"[G]ood academic English is often described as *error-free*" (McKinley & Rose 2018: 1). However, before discussing what constitutes an *error-free text*, it is important to define what an error[1] is. The present study uses a definition taken from the field of Applied Linguistics, which defines an error as "a linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native speaker counterparts" (Lennon 1991: 182). This definition seems fitting as McKinley and Rose's (2018: 9) analysis of journal guidelines shows that expected language norms were often benchmarked "elusively against a native-speaker yardstick, which was not clearly defined" and that L2 English writers were commonly advised to "have their language checked 'by a native speaker', [which] positioned native speakers as authorities on error-free writing".

However, Lennon (1991: 185) mentions that "the appeal to native speaker intuitions alone was found to be inadequate as a criterion for error" and "that only by setting up […] a corpus-specific error definition, tied to a reference group, mode and situation, and probabilistic rather than deterministic, could attempts be made to identify error" (185). Accordingly, in the present study errors will not be judged from a prescriptivist position. Instead, errors will be judged from a corpus-linguistic standpoint meaning that only instances of language use that do not or only very rarely occur in native speaker corpora[2] will be judged as errors. This allows for conclusions to be drawn that are more reliable and representative than the intuition of a single native speaker.

[1] Errors should not be confused with so-called mistakes, which are "'the product of "chance circumstances", analogous to such native speaker slips of the tongue' (Corder 1967:166)" (Lennon 1991: 181).

[2] Liu and Lei (2017) define a corpus as "a balanced and principled collection of naturally occurring (i.e., occurring in real communication situations) written texts or transcripts of audio/video recordings compiled to represent a certain language or language variety" (2).

From an empirical perspective, various studies have found a corpus-based approach to academic writing and error correction to be quite effective, especially in the ESL and EFL context (e.g. Quinn 2015; Bridle 2019; Dolgova & Mueller 2019; Flowerdew 2022). Working with a corpus can be seen as a form of empowerment for non-native speakers (Mair 2002) because they can use corpora as a source of authentic data (e.g. Bridle 2019) to help them solve language issues and, thereby, improve their writing quality (e.g. Quinn 2015).

2.3 AI Tools as the miracle cure?

"Accurate language use, especially in written language, is part of the academic message" (Turner 2004: 108) and it is good practice for a paper to "be proof read before sending off to the journal for review" (Gilmore et al. 2006: 470). Thus, the arrival of new AI writing tools might seem like the ideal solution: both time-efficient and inexpensive.

The term *Artificial Intelligence* is used to describe "the ability of computer systems to mimic human intelligence" (Bowen & Watson 2024: 16). These systems are trained for specific purposes, derive their "knowledge" from millions of examples and use heuristics and algorithms to make predictions (Rich & Knight 1991). *Generative Artificial Intelligence* "is any type of AI that uses deep learning models to generate or create new content" (Bowen & Watson 2024: 22). Bowen & Watson (2024) adduce *Large Language Models (LLMs)*, which "use a combination of technologies to mimic human language and predict the next word" (17), as one example. However, the term *Artificial Intelligence* is misleading (Clarke 2023): these models try to imitate human intelligence (Buck 2025: 24), but they are not (yet?) intelligent per se.

There are many AI applications that can be used for educational and academic purposes. Current research in this field aims to investigate potential benefits and limitations (e.g. Imran & Almusharraf 2023; Lo 2023; Warschauer et al. 2023; Leitgeb & Leitgeb 2024) as well as applications (e.g. Cummings et al. 2024; Koltovskaia et al. 2024; Hwang et al. 2025; Tan 2025; J. Xu 2025) in an academic teaching and writing context. One study revealed that across all investigated disciplines more than 50% of the German students who took part in the survey had used AI tools regularly during the winter term 2024/25 (Hüsch et al. 2025). Furthermore, specific guides on how AI can be leveraged as aids for academic writing (e.g. Buck 2025) or for education in general (e.g. Bowen & Watson 2024; Kohnke 2026; Son & Philpott 2026) have already been published.

Studies are currently emerging that investigate how AI tools are being used by university students for academic writing purposes. In their scoping review, Bo et al. (2025) review 20 previous empirical studies on AI usage in higher education. They conclude that most previous studies focused on how AI tools can be used for improving students' academic writing and that these "studies reported improvements in grammatical accuracy, vocabulary use, text cohesion, and

argument structure when AI-assisted tools were used" (Bo et al. 2025: 104). The authors also find that most of the previous studies were carried out in EFL contexts. To this author's knowledge, no studies have been conducted in an L1 context so far, so as of now no conclusions can be drawn regarding whether the use of AI tools for academic writing differs between native and non-native users of English.

Cengiz et al. (2025) investigate *ChatGPT*'s effectiveness as a writing feedback tool based on students' revision behaviours and perceptions. The authors conclude that

> [t]he high acceptance rate of language use feedback and the frequent use of direct substitutions from ChatGPT, though partly rooted in low language proficiency, suggest a risk of over-reliance on AI-generated corrections and reduced critical thinking during the revision process (Cengiz et al. 2025: 10).

Their observations highlight the need for *AI literacy* (D. T. K. Ng et al. 2021; OECD 2025). This term encompasses future skills that must be acquired in order to evaluate AI tools instead of merely using them in an unreflective way. For example, little is known about the training data used to develop the language models that AI tools are based on[3] – are they even representative of (academic) English? It remains a *desideratum* to check whether corrections provided by these tools are indeed adequate in a specific context.

There is little empirical research that analyses the use of AI tools for error correction purposes with regard to how AI tools perform and most of this research has a computer science focus (e.g. Coyne et al. 2023; Fang et al. 2023; Wu et al. 2023) rather than a linguistic one. One conclusion that can be drawn from previous research in this area is that different versions of *ChatGPT* have varying performance rates (Coyne et al. 2023; Schlüter & Deckert 2025). Furthermore, *ChatGPT* shows varying correction performances depending on the error types (Fang et al. 2023; Lechien et al. 2023) and its performance and correction rates decrease with increasing sentence difficulty and length (Fang et al. 2023; Wu et al. 2023; Y. Xu et al. 2024). Most studies reveal that *ChatGPT* tends to generate over-corrections, preferring more extensive or even superfluous changes to minimal corrections (Coyne et al. 2023; Fang et al. 2023; Lechien et al. 2023; Wu et al. 2023), but it only rarely mis-corrects and, therefore, usually manages to maintain grammatical correctness (Wu et al. 2023). Moreover, *ChatGPT* seems to perform better in high-level writing compared to low-level writing (Pfau et al. 2023). It has been pointed out that *ChatGPT* requires a suitable and precise prompt in order to function correctly (Coyne et al. 2023; Lechien et al. 2023; Y. Xu et al. 2024), and that temperature[4] settings (Coyne et al. 2023; Y. Xu et al. 2024) and techniques

[3] There is one paper (Brown et al. 2020: 7) that describes the composition of *GPT-3*'s training data, which reveals that most of the data was collected from different archives and corpora containing web data, books and Wikipedia.

[4] In this context, temperature is "a parameter that controls the randomness of the response generated by ChatGPT" (Ouyang et al. 2025: 2).

such as Chain-of-Thought (CoT) prompting (Fang et al. 2023) can influence correction performance. While *ChatGPT* seems to have "some potential for a relatively quick measure of linguistic accuracy for research purposes" (Pfau et al. 2023: 7), human intervention and evaluation is still required to verify the output (Pfau et al. 2023; Xu et al. 2024).

Buck (2025: 212–213) supposes that AI tools specifically dedicated to error correction (e.g. *DeepL Write*, *LanguageTool*) perform better on error correction tasks than multi-purpose AI tools (e.g. *ChatGPT*, *Claude*). Most prior research only investigates the all-rounder *ChatGPT* while specifically designed text correction tools are only rarely in focus. One study analyses *LanguageTool* and concludes that "the errors it locates are generally accurate, but it fails to detect many errors" (Crossley et al. 2019: 265). Wu et al. (2023) and Fang et al. (2023) compare *ChatGPT* and *Grammarly*: while the former find *Grammarly* to perform better and more stably in general, the latter note that for fluency corrections *Grammarly* performs better on shorter sentences, while *ChatGPT* performs well for all sentence lengths. Schlüter and Deckert (2025) compare three different AI tools – *DeepL Write*, *ChatGPT* (*GPT3.5* and *GPT4*) and *Grammarly* – and find that *DeepL Write* performs best whereas *Grammarly* performs worst. To the author's knowledge, the latter are the only three studies that directly compare dedicated and multi-purpose AI tools regarding their error correction performance. Since by now there is a plethora of freely available dedicated and multi-purpose AI tools online, further research is required to determine which (type of) tool performs best in which circumstances.

This study seeks to advance prior research by adding a linguistic focus, i.e. classifying errors not merely based on prescriptive rules (which AI tools are expected to "know") but instead by classifying an error as such based on data depicting authentic native speaker usage taken from language corpora. The errors were selected to be typical for German learners of English, meaning that conclusions drawn from this study are mainly applicable for a German EFL context. Nevertheless, the results can be interesting for learners from different language backgrounds as well because the error types chosen for analysis (prepositional usage, collocations, easily confusable words, grammatical structures and academic register) do not exclusively occur with German learners.

Furthermore, in this study not only one single AI tool will be used to correct the errors. Instead, a selection of AI writing tools will be employed. This study does not use a random selection of errors but specifically chooses a number of errors from different problem areas that learners of English often struggle with in order to evaluate whether these tools perform better for specific error types. Finally, two different evaluation metrics (*detection accuracy* and *correction suitability*) will be used to assess the tools' performances.

2.4 Research questions

AI tools can help (non-)native students and researchers from various fields of study using English as a *lingua franca* by correcting their academic texts. A thorough linguistic evaluation of AI corrections based on corpus data to judge their *detection accuracy* and *correction suitability* is necessary to competently and confidently employ the tools as *artificial proofreaders*. The paper at hand aims to provide an exploratory analysis to answer the following research questions:

RQ1: How effective are the AI tools under scrutiny for error correction purposes from a corpus linguistic standpoint?
RQ2: Do dedicated AI writing tools perform better at error correction than a multi-purpose AI tool?
RQ3: Do different error types lead to different performance rates?

3 Methods

3.1 Generation of fictitious learner texts

The present study builds on a pilot study conducted by Schlüter and Deckert (2025), which already highlighted varying performance rates between three AI tools: *ChatGPT*, *Grammarly* and *DeepL Write*. As a starting point for the present study, *GPT4o-mini* was employed to generate[5] an academic text such as could have been written by a German non-native student of English on an essay topic that was chosen randomly (in this case, tools for improving writing skills) with the aim of having a text (T1; see Appendix, Figure 2) that could be used for an evaluation of AI tools and their correction performance. The fictitious learner text was further manipulated by manually inserting 50 typical German learner errors[6], with ten belonging to each of five problem areas: prepositions, collocations, confusable words, grammar and register. This resulted in a text such as could have been written by an advanced German EFL learner at university level.

The decision to have the errors included in a text rather than having them corrected by AI tools without context was deliberate to make the analysis as

[5] This was the original prompt used to generate the first draft of the learner text: `I am a university student in my third semester studying English for a Bachelor's degree. My mother tongue is German. I have to write an academic paper for a seminar. The topic is supposed to be "Out with corpora, in with ChatGPT & Co.? Using corpora and AI to improve your writing skills". Can you write an introduction for my paper?` The text resulting from this prompt was rephrased (sentences added, sentences restructured, vocabulary changed) to include certain phrases and structures that might be problematic for German learners of English and are, therefore, prone to errors.

[6] These errors were considered as typical based on previous literature (e.g. Alexander 1994; Fitikides 2002; Parkes 2003) as well as the author's own experience as a teacher of English to German university students.

authentic as possible. When students and scholars employ these tools, they usually will not be aware of the errors they made but will simply paste longer text passages into the AI tools to both detect and correct errors. The decision to employ a fictitious learner text rather than authentic learner texts was also deliberate in order to include a larger number of specifically chosen errors. Control of the types and number of errors thus comes at the expense of authenticity, which must be kept in mind when discussing the results.

Two more texts including the same errors in a different order and in varying semantic contexts were generated[7]. For these additional texts "medicine" (T2; see Appendix, Figure 3) and "legal education" (T3; see Appendix, Figure 4) were chosen as the respective topics. The different fields only served the purpose of having three texts that differed in terms of the topics while still including the same errors. Conclusions on whether there is a difference in error detection and correction between varying academic disciplines cannot be deduced from this analysis.

Since errors were defined as instances of language use that do not or only very rarely occur in native speaker writing, a corpus linguistic approach was used to identify the errors as such. The author employed three corpora for manual error identification: the *Corpus of Contemporary American English* (COCA; Davies 2008-), the *British National Corpus* (BNC; Davies 2004) and the *Global Web-based English* (GloWbE; Davies 2013) corpus. All three corpora were chosen because they represent present-day English in the two main varieties (COCA and BNC) as well as numerous World Englishes (GloWbE) and include texts in an academic register (COCA and BNC).

3.2 Exemplary errors from the problem areas

The three texts including the errors, their adequate corrections, their error code and their allocation to one of the five problem areas can be found in the Appendix. It is important to note that errors made by advanced learners of English differ from errors made by beginners: "At the advanced level highly localized morphological error is less prevalent than error involving usage, lexical choice, stylistic appropriacy, and various sorts of global discourse error" (Lennon 1991: 184). Therefore, these errors are more difficult to spot.

[7] For the additional learner texts, a different prompt was used: `Can you give me a text that is about 400 words long and serves as an introduction for an academic term paper written for a seminar? It is supposed to be written by a university student in their third semester studying medicine/legal education. Their mother tongue is German. The following words and phrases need to be in the text: […].` The goal was to generate two more texts that would include the same words and phrases that were used as errors in the first text. To achieve this, the AI tool was instructed to include the correct words and phrases first. Afterwards, the texts were once again manipulated by manually replacing the correct words and phrases with the 50 typical German learner errors.

For the purposes of this study, advanced learner errors were chosen as it was assumed that beginners' errors would be too easily solved by AI tools. Information on and examples of advanced learner errors are harder to find in the literature. Therefore, different works of previous research on common learner errors (e.g. Alexander 1994; Fitikides 2002; Parkes 2003) were consulted for typical types of learner errors. These errors were then adapted to fit both an advanced and a German learner context based on the author's experience.

In the following sub-sections, the errors used in the texts will be listed along with an explanation of why they are considered typical learner errors, where applicable accompanied by a reference to previous research. One detailed example for one error from each of the problem areas as well as an explanation on how it was classified as such using the corpora will be provided in the following five sub-sections. All errors were identified as such using three corpora: the COCA, the BNC and the GloWbE. For simplicity and brevity, the identification process in the following sub-sections will be illustrated using only the COCA as an exemplary corpus. A complete overview of all errors along with their assigned problem area, their error code for future reference and their model correction can be found in the error glossary (see Appendix, Table 3).

3.2.1 Prepositional usage

Prepositional usage is an area in which many learners from different language background experience difficulties as it is prone to L1 interference (Rankin & Schiftner 2011) as "[a]lmost every one of the usage rules for […] prepositions has exceptions or subrules, and they interact in complex ways, which makes it much harder for English language learners to acquire the correct usage" (Leacock et al. 2014: 28). Since prepositions pose such severe problems for learners, they are categorised as a problem area of their own in this paper, rather than as merely one sub-type of grammatical or collocational errors.

For example, the German equivalent of the noun *majority* ('Mehrheit') takes the preposition *von*, which is often translated into English as *from*. Learners might assume that it should then be *majority from* in English (PR3; e.g. T2: "the majority from healthcare practices"). A corpus search[8] shows that the most frequently used preposition after the noun *majority* is *of* (48,566 hits) compared to only 83 hits for *majority from*. The preposition *of* is used in more than 90% of all instances of prepositions following the noun *majority*, while *from* is used in not even 1% of all hits. Therefore, even though it occurs in the corpus, *majority from* should not be considered as a variant but as an error.

The most common prepositional errors mentioned in the error guides consulted, for instance "depend from" (Fitikides 2002: 4; Parkes 2003: 125) or "good in" (Fitikides 2002: 7; Parkes 2003: 125), are not as problematic for most

[8] Link to corpus search: https://www.english-corpora.org/coca/?c=coca&q=126688677

advanced learners, while less common prepositions remain difficult. Prepositional errors were chosen either because the use of the faulty preposition could very likely be the result of L1 interference: "emergence from" (PR5), "exemplified through" (PR6), "potential from" (PR8), "through exploring (PR9) and "underpinnings from" (PR10); or because the faulty and the correct prepositions were very similar and, therefore, easily confusable: "has led into" (PR1), "revolution at" (PR2), "INTEGRATE … to" (PR4), "approach at" (PR7).

3.2.2 Collocations

Collocations are "the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance" (Gries 2008: 6). Which words frequently co-occur apparently has "no rules and therefore must be memorized" (Leacock et al. 2014: 28), making it challenging to find the correct collocation for a specific context, especially for learners of a language.

When referring to new or up-and-coming technologies, learners might be unsure which descriptive adjective to use (CO6; e.g. T1: "surfacing technologies"). A search for adjective collocates of the noun *technology*[9] reveals that the most frequently used adjective collocate with a meaning equivalent to the German *aufkommend* is *emerging* (807 hits). *Surfacing* does not appear among the 200 top collocates of *technology*.

Common collocational errors, for instance "make homework" or "hold a speech", are not as problematic for most advanced learners, while less common collocations remain difficult. This is why the collocational errors used in the texts were not chosen based on the literature but due to possible L1 interference: "priceless insights" (CO1), "wealthy repository" (CO3), "elementary part" (CO3), "ENHANCE the importance (CO4), "running dialogue" (CO5), "deep implications" (CO7), "FACILITATE learners" (CO8), "meaningful element" (CO9), "insights it supplies" (CO10; e.g. T1).

3.2.3 Easily confusable words

Confusable words (e.g. homonyms, near-synonyms, converses, false friends) are words that cannot be used interchangeably but resemble each other in terms of form and/or meaning and are, therefore, easily confused by learners of a language (Liu 2013b). The English noun *concurrence*, meaning an agreement between two parties or two things happening at the same time, and its German cognate *Konkurrenz*, correctly translated as *competition* (Alexander 1994: 39), are one example of false friends (CW9; e.g. T1: "In navigating the concurrence of traditional and surfacing

[9] Link to corpus search: https://www.english-corpora.org/coca/?c=coca&q=131928918

technologies"). To differentiate confusable nouns, their co-occurring determiners, adjectives, post-nominal infinitive modifiers or verbs can be analysed (Liu 2013a). For this noun pair, their verb collocates offer the most revealing insights. A corpus search shows that one can *face*, *eliminate* and *reduce* the *competition* (832, 192 and 152 hits, respectively), but one can neither *face*, nor *eliminate*, nor *reduce* the *concurrence*[10] (0 hits).

Other confusable words used in the texts can also be distinguished in corpora by comparing their most common collocates, such as *approximate* and *approach* (CW2), *individuate* and *individualize*/*individualise* (CW4), *intrigues* and *intricacies* (CW5), *compliment* and *complement* (CW6; cf. Alexander 1994: 39), *significance* and *importance* (CW7), *lever* and *leverage* (CW8) as well as *appliances* and *applications* (CW10). Two other word pairs can be distinguished by their frequency differences as the faulty word does not occur in the given context at all: *digitalized*/*digitalised* and *digital* (CW1) as well as *procure* and *provide* (CW3).

3.2.4 Grammatical structures

One example of problems occurring in the field of grammar is the *horror aequi* effect (GR3; e.g. T2: "educators are beginning integrating new technologies"), which is thought to be a cognitive mechanism that makes speakers "avoid repetitions of similar (or even identical) linguistic structures operating on various levels of linguistic organization" (Baumann & Mühlenbernd 2022: 51). For instance, while the verb *begin* can take gerunds or *to*-infinitives as complements, adjacent *-ing* forms in structures such as the progressive form *beginning* followed by a non-finite verb ending in *-ing* tend to be avoided (Rohdenburg 2003). The fact that two successive *-ing* forms (only 14 hits with *beginning* as the first element) seem to be avoided can also easily be shown with a corpus search[11]. In comparison, *beginning* followed by a *to*-infinitive[12] (19,855 hits) is a very frequent grammatical construction.

The *horror aequi* phenomenon is only one instance in which the use of a gerund would be deviant. Fitikides (2002: 17) also remarked on the gerund as a source of errors, which is why several such errors were included in the learner texts: "promise redefining" (GR1) "tendency of improving" (GR5), "aims at shedding light" (GR6), "ENABLE … at achieving" (GR9). The remaining grammatical errors concern agreement with collective nouns in "the majority of … VERB" (GR2), the misuse of the present continuous (Fitikides 2002: 23) in "this paper is exploring" (GR4), the use of unnecessary prepositions (Fitikides 2002: 61; Parkes 2003: 125) in "discuss about" (GR7), the omission of a preposition (Fitikides 2002:

[10] Link to corpus search: https://www.english-corpora.org/coca/?c=coca&q=125929490
[11] Link to corpus search: https://www.english-corpora.org/coca/?c=coca&q=121063706
[12] Link to corpus search: https://www.english-corpora.org/coca/?c=coca&q=121358357

46) in "the question how" (GR8) and the misplacement of an adverb (Fitikides 2002: 75; Parkes 2003: 121) in "is enough indicative" (GR10).

3.2.5 Academic register

Staples et al. (2015: 505) define registers as "named varieties within a culture" that "can be defined at any level of generality" and state that "there are systematic and important linguistic differences across registers" (506). However, the mastery of these differences is "not inevitable or even easy for either the native or the foreign learner of a language" (Quirk et al. 1985: 26) and can, therefore, pose problems. The register targeted in this study is the academic style employed in research papers. Since both the COCA and the BNC are composed of different registers (including academic writing), these corpora can be used to evaluate whether a certain word, phrase or structure is appropriate in an academic text. With regard to register issues, it is important to note that the use of a certain word or phrase is often not erroneous in general and can be entirely appropriate in one register but inappropriate in another.

There are different ways of expressing the passive in English: the *be*-passive (e.g. T3: "it has to be kept in mind"), and the *get*-passive (e.g. T3: "ethical considerations should not get overlooked"). Corpus searches for the *be*-passive[13] and the *get*-passive[14] reveal that both structures are frequently used and can, therefore, be judged as grammatically correct. The *be*-passive is used most frequently in the academic register (1,121,821 hits), while the *get*-passive (RE2; RE4) is characteristic of informal and spoken language and very rare in academic texts (1,739 hits). When writing in an academic context, the *be*-passive is the better choice.

Other register errors used in the texts belong to the area of contractions ("there's" (RE5) and "shouldn't" (RE7)), lexical choices ("a lot of" (RE1), sentence-initial "But" (RE3), "gigantic" (RE6), "pretty innovative" (RE8), "MAKE … better" (RE10)) and the use of personal pronouns ("your" (RE9)). All of these errors can easily be identified as such by conducting a CHART search in the corpora.

3.3 Correction process with AI tools

The final version of T1 was corrected by nine AI tools[15]. Two different approaches were used depending on the AI tools: For the eight dedicated writing tools (*DeepL Write, Grammarly, Trinka.ai, LanguageTool, ProWritingAid, Wortliga, Paperpal,*

[13] Link to corpus search: https://www.english-corpora.org/coca/?c=coca&q=121226902
[14] Link to corpus search: https://www.english-corpora.org/coca/?c=coca&q=119750081
[15] Many of these tools offer users subscriptions and improved versions for a fee. In this study, only the freely available versions of the tools were tested because these are the versions available to everyone.

and *Mimir Mentor*) two task versions were conducted: one for the text in British English (BrE; including stereotypically British spellings, e.g. *recognise, personalise*) and another one for the text in American English (AmE; including stereotypically American spellings, e.g. *recognize, personalize*). If possible, the target variety was set before correction, and further settings were adjusted in accordance with the context envisaged for the text: a formal academic writing style. The only exception was *DeepL Write*, as choosing an academic writing style did not only lead to corrections but to entire passages of the text being rephrased. Therefore, a neutral writing style was chosen for *DeepL Write*. *Trinka.ai* is available in two different modes: lite and power. Since choosing the power mode also led to entire passages of the text being rephrased and, therefore, to error contexts that were to come under scrutiny being deleted, the lite mode was used. The eight dedicated writing tools were selected because they are either commonly known, had been investigated in previous research or are freely available and, therefore, quite enticing to use.

For the multi-purpose AI tool *ChatGPT* (*GPT4o-mini*[16] and *GPT5*[17]), a prompt had to be designed:

```
Can you help me correct the following text written by a non-native
writer who is a university student of English? It should be written
in [American/British] English for an informed, academic audience.
Could you focus on making corrections only, without rephrasing too
much of the sentence? And could you highlight the corrections you
made in your text output in bold print?
```

The prompt was created based on Giray's (2023) guide for prompt engineering.

Afterwards, the error-containing text was inserted into the chat. Since *ChatGPT*'s output is probabilistic[18] and, therefore, inconsistent across attempts (Pfau et al. 2023; Y. Xu et al. 2024), a total of 100 iterations were conducted to gauge the variance of the correction performance, with 50 iterations correcting the British version and 50 iterations correcting the American version of the text.

It needs to be pointed out that using two task versions for dedicated AI writing tools compared to 100 iterations for the multi-purpose AI tool *ChatGPT* is necessary for this evaluation as dedicated AI writing tools, compared to generative AI tools, are not probabilistic and do not offer different solutions during each correction. Therefore, conducting only two runs is entirely sufficient. Generative AI tools, however, have this probabilistic property and provide a different answer

[16] In the present study, the *HAWKI* version (hawki.its.uni-bamberg.de) of *ChatGPT* was used, which is a didactic interface for universities based on the OpenAI API. In the version offered by the University of Bamberg *GPT4o-mini* is used as a base model.
[17] To test the newest language model *GPT5*, OpenAI's *ChatGPT* interface (chatgpt.com) was used.
[18] There are ways of making *ChatGPT*'s output less probabilistic and more deterministic. When creating custom *GPT*s with an API, creators can choose to change the temperature settings. However, *ChatGPT*'s freely available default version and the *HAWKI* version do not offer this option.

each time they are given a prompt. Therefore, it is necessary to conduct many runs (100 in this case) to statistically evaluate their full potential.

3.4 Evaluation metrics

Error corrections were judged from two perspectives: A *detection accuracy* score ($F_{0.5}$) was used to judge how accurately tools detected errors. Additionally, a *correction suitability* score (percentages) was used to evaluate whether the corrections were adequate.

Four codes were used to evaluate the tools' (non-)corrections: "error not found", "adequate correction", "inadequate correction" and "error deleted". A fifth code was added because *ProWritingAid* and *Grammarly* sometimes correctly marked words as errors but would only provide the correct version with the premium license and *Trinka.ai* occasionally indicated an error and gave explanations for why a word was not suitable but without providing a correction. These instances were then coded as "unfinished correction".

Furthermore, the AI tools did not only correct the target errors that were deliberately inserted into the texts but also made additional changes. Although additional changes are part of the *detection accuracy* score, these changes will not be analysed qualitatively in further detail in this paper[19].

3.4.1 *Detection accuracy* score

$F_{0.5}$ scores were calculated as an evaluation metric for the tools' *detection accuracy*. This metric can be employed to measure accuracy by combining precision and recall (H. T. Ng et al. 2014). Precision was calculated by dividing *True Positives* (i.e. instances in which a tool accurately detected an error) by the sum of *True Positives* and *False Positives* (i.e. instances in which a tool falsely identified something as an error that was actually correct). Recall was calculated by dividing *True Positives* by the sum of *True Positives* and *False Negatives* (i.e. instances in which a tool failed to correct an actual error). $F_{0.5}$ scores were chosen because they are commonly used in Grammatical Error Correction Tasks as "$F_{0.5}$ emphasizes precision twice as much as recall" (H. T. Ng et al. 2014: 6) and because "[w]hen a grammar checker is put into actual use, it is important that its proposed corrections are highly accurate in order to gain user acceptance. Neglecting to propose a correction is not as bad as proposing an erroneous correction." (H. T. Ng et al. 2014: 6).

[19] Although additional changes are not the focus of this paper, at this point, it must be noted that they are not always changes for the better. They can either be used interchangeably with the original version, or they might even introduce new errors in the text (e.g. by distorting the meaning or omitting information). Regardless of their appropriacy, however, they can also cause insecurity in users, who might be led to believe that their original expression was an error and actually in need of correction.

For the present study, corrections and changes assigned to the categories "adequate correction", "error deleted" "inadequate correction" and "unfinished correction" were considered as *True Positives* because in these instances the errors were detected by a tool. Additional changes were considered as *False Positives* because a tool classified something as an error which was not in need of correction and cases of "error not found" were considered as *False Negatives* because tools failed to spot an error in such instances.

3.4.2 *Correction suitability* score

Percentages were used as an evaluation metric for the tools' *correction suitability* with the goal of evaluating whether a correction that had occurred was suitable. Adequate corrections were judged as *target corrections*, since the original error was amended. Inadequate corrections and instances in which an error was not found at all or changed in a non-target way were evaluated as *no or off-target corrections*, since an error remained in the text. Unfinished corrections were not included in either category but were judged separately as they did not immediately provide a possible correction. Error deletions were evaluated more closely. Depending on whether the error deletions were judged as adequate or inadequate, they counted towards either *target corrections* or *no or off-target corrections*.

4 Results

4.1 RQ1: How effective are the AI tools under scrutiny for error correction purposes from a corpus linguistic standpoint?

A preliminary analysis of T1 was conducted with all nine tools. Only tools that either made 45% target corrections (*correction suitability*) or had an $F_{0.5} > .70$ (*detection accuracy*) were chosen for further analysis. Table 1 shows that there are clear differences between the tools regarding the evaluation metrics: *DeepL Write (default)* scored highest for *correction suitability*, while both *GPT5* and *GPT4o-mini* had the highest $F_{0.5}$ score.

Table 1: Overall performance (*correction suitability*) and $F_{0.5}$ scores (*detection accuracy*) of AI tools' error corrections for T1.

Tools	Target corrections	No or off-target corrections	Unfinished corrections	$F_{0.5}$
DeepL Write[a] (default)	82%	18%	0%	.66
GPT5[b]	68%	32%	0%	.93
GPT4o-mini[b]	64%	36%	0%	.93
Paperpal[a]	62%	38%	0%	.85
Wortliga[a]	59%	41%	0%	.91
Grammarly[a]	45%	25%	30%	.88
Trinka.ai[a] (lite mode)	43%	55%	2%	.81
DeepL Write[a] (corrections only)	37%	63%	0%	.73
ProWritingAid[a]	24%	48%	28%	.76
LanguageTool[a, c]	23%	77%	0%	.59
Mimir Mentor[a]	6%	94%	0%	.27

Notes: [a] Two task versions (one for BrE and one for AmE) were conducted for the dedicated AI writing tools. [b] 100 iterations (50 for BrE and 50 for AmE) were conducted for both *GPT*s. [c] For *LanguageTool* it is also important to note that hardly any corrections were made in the text when set to British English. More corrections were made when set to American English.

The correction performances of *LanguageTool* and *Mimir Mentor* remained below both thresholds, so these tools were not analysed any further. The other tools made at least 45% target corrections or scored above .70 for their $F_{0.5}$ score and were, therefore, analysed for RQ2.

4.2 RQ2: Do dedicated AI writing tools perform better at error correction than a multi-purpose AI tool?

4.2.1 *Detection accuracy* ratings

Calculating $F_{0.5}$ across all three texts (Table 2) indicates discrepancies regarding *detection accuracy* between the different tools. *GPT5*, one version of the multi-purpose AI tool, achieved the highest mean $F_{0.5}$ score, closely followed by *Wortliga*

and the older *GPT4o-mini*. *Grammarly*, *Paperpal*, *Trinka.ai*, *DeepL Write (corrections only)* and *ProWritingAid* only achieved a mediocre score. *DeepL Write (default)* scored even lower, below the .70 threshold. For this analysis and the errors and texts at hand, there seems to be a distinction between the dedicated AI tools and the multi-purpose AI tool regarding their *detection accuracy* as almost all dedicated AI writing tools – except for *Wortliga* – have a lower $F_{0.5}$ score compared to the two *GPT*s.

4.2.2 *Correction suitability* ratings

The contrastive analysis across all three texts (Table 2) indicates notable differences regarding the tools' capabilities for *correction suitability*, with *DeepL Write (default)* achieving the highest overall performance at 76% target corrections. For the overall comparison, *ProWritingAid* performed worst with 25% target corrections. *Correction suitability* varied slightly across texts.

One dedicated AI writing tool, *DeepL Write (default)*, had a very high rate for *correction suitability*, others (*Paperpal, Trinka.ai* and *Grammarly*) a rather low one. Neither *GPT* version achieved the highest or the lowest error correction rate, but the newer version *GPT5* still outperformed the older *4o* version. The tools that remained below the threshold (< 45% target corrections) were dedicated AI writing tools. Thus, for this analysis and the errors at hand there seems to be no clear distinction between the dedicated AI tools and the multi-purpose AI tool regarding their *correction suitability*.

Table 2: Overall performance (*correction suitability*) and mean $F_{0.5}$ scores (*detection accuracy*) of AI tools' error corrections across the three texts.

Tools	Target corrections			No or off-target corrections				Mean unfinished corrections	Mean $F_{0.5}$ score
	adequate corrections	adequate deletions	**Mean**	errors not found	inadequate corrections	inadequate deletions	**Mean**		
DeepL Write[a] (default)	64.3%	11.7%	**76%**	1.3%	7.3%	15.3%	**24%**	**0%**	**.61**
GPT5[b]	69.4%	0.0%	**69%**	20.4%	8.4%	1.9%	**31%**	**0%**	**.91**
GPT4o-mini[b]	60.0%	0.0%	**60%**	29.6%	7.4%	3.0%	**40%**	**0%**	**.89**
Wortliga[a]	57.0%	0.0%	**57%**	33.7%	8.3%	1.0%	**43%**	**0%**	**.90**
Paperpal[a]	51.0%	0.3%	**51%**	41.3%	2.0%	5.3%	**49%**	**0%**	**.78**
Trinka.ai[a] (lite mode)	45.3%	0.0%	**45%**	47.7%	1.7%	3.3%	**53%**	**2%**	**.78**
Grammarly[a]	46.3%	0.0%	**46%**	23.0%	2.0%	1.0%	**26%**	**28%**	**.79**
DeepL Write[a] (corrections only)	37.0%	0.7%	**38%**	59.7%	2.0%	0.7%	**62%**	**0%**	**.76**
ProWritingAid[a]	25.3%	0.0%	**25%**	46.7%	2.7%	0.0%	**49%**	**25%**	**.72**

Notes: Percentages may not sum to 100% due to rounding. [a] Two task versions (one for BrE and one for AmE) were conducted for the dedicated AI writing tools. [b] 100 iterations (50 for BrE and 50 for AmE) were conducted for both *GPT*s.

4.2.3 A qualitative analysis of inadequate corrections and error deletions

In the following, two types of changes AI tools made will be analysed in more detail, namely those instances in which a tool neither corrected an error in the predicted manner nor missed it entirely but rather altered an error in a way that was not expected: inadequate corrections and (adequate and inadequate) error deletions. An overview of the different instances of inadequate changes and error deletions can be found in the error glossary (see Appendix; Table 3).

In total, *DeepL Write (default)* made the most inadequate changes and the number of inadequate changes this tool made even exceeded the number of errors the tool missed entirely. *GPT5*, *GPT4o-mini* and *Wortliga* also produced quite a substantial number of inadequate corrections. There are two reasons why inadequate corrections were considered as such: Some inadequate corrections changed the original meaning of the text: for example, some tools replaced the confusable word *individuate*, which should have been corrected to *individualise*, with the verbs *supplement*, *augment* or *enhance* (CW4). Other inadequate corrections were judged as grammatically incorrect: for instance, some tools replaced the preposition *at* after *approach*, which should have been corrected to *to*, with the preposition *in* (PR7). *DeepL Write (default)* only made inadequate corrections that changed the original meaning. Most of the inadequate corrections *Wortliga, GPT4o-mini* and *GPT5* made also changed the original meaning, while some others were grammatically incorrect.

For almost all tools, error deletions occurred only very rarely and contributed less than 5% of their total error corrections. The only exceptions are *DeepL Write (default)*, which made by far the most error deletions, and *Paperpal*.

When analysing instances of error deletion, it must be noted that there are two manners in which tools deleted errors. In some cases, only the erroneous words or phrases themself were deleted: for example, the modifier *pretty* in "pretty innovative" (RE8). In other cases, entire sentences were rephrased, which caused erroneous words or phrases to be deleted: for instance, the phrasal verb *lead into* was replaced by the single verb *sparked* (PR1). Most of *DeepL Write (default)*'s error deletions occurred because of extensive rephrasing, while *Paperpal* mostly deleted erroneous words or phrases.

Furthermore, error deletions can either be judged as adequate if they did not alter the meaning of the text (e.g. through shortening, syntactic simplification or changed wording) or as inadequate if they led to a loss of information in the text or changed the original meaning. In total, both *DeepL Write (default)* and *Paperpal* made more inadequate deletions compared to adequate deletions.

4.3 RQ3: Do different error types lead to different performance rates?

4.3.1 *Correction suitability* ratings

The *correction suitability* differed not only depending on the AI tool under scrutiny but also depending on the error type that was corrected. Before looking at the results, it should be kept in mind that the learner texts only include a limited selection of errors. This was deemed necessary to produce a small and controlled dataset that can be analysed in detail. However, this comes at the expense of broader generalisations based on the results and the results should be seen as a first foray into potential differences according to error types.

Only tools that made at least 45% overall target corrections were considered for this analysis as tools that scored very low for overall *correction suitability* would score very low in all error categories and, thereby, distort the results. Figure 1 shows that across all texts, tools scored at least 70% target corrections for grammar problems and at least 60% target corrections for prepositional errors. Overall, all tools make more target corrections for prepositional and grammatical errors compared to the other problem areas. Correction performance was mixed for confusable words, with some tools (*DeepL Write (default), GPT4o-mini, GPT5*, *Wortliga* and *Grammarly*) performing rather well (> 50%) and other tools (*Paperpal* and *Trinka.ai*) performing worse. Compared to the other problem areas, correction performance for register errors varied considerably with *Paperpal*, *Trinka.ai* and *DeepL Write (default)* achieving higher scores, while *GPT4o-mini*, *GPT5*, *Wortliga* and *Grammarly* did not catch many register errors. Collocations seemed to be difficult errors for almost all tools to spot and amend correctly – with the exception of *DeepL Write (default)* and *GPT5*. *DeepL Write (default)* stands out as it makes at least 60% target corrections in each problem area, even though for prepositional errors both *GPT5* and *GPT4o-mini* perform slightly better.

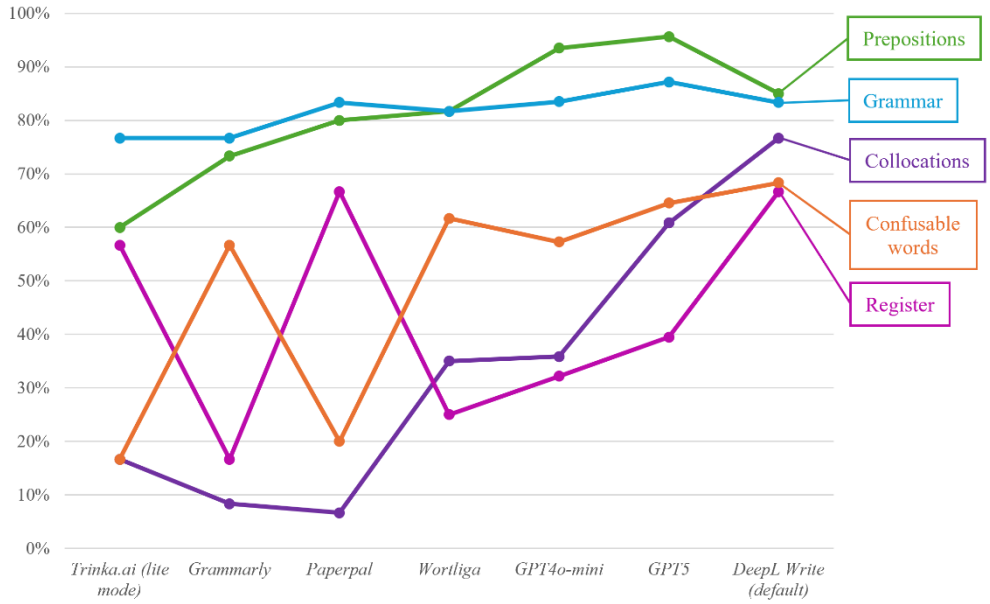

Figure 1: Mean performance (*correction suitability*) of AI tools' error corrections for the five problem areas across the three texts.

4.3.2 A qualitative analysis of inadequate corrections and error deletions

Most inadequate changes that altered the original meaning of the text belong to the problem area confusable words, e.g. the noun *concurrence*, which should have been corrected to *competition* or *rivalry*, was changed to *convergence*, *adoption*, *advancement* or *intersection* (CW9). Some other inadequate changes occurred with collocational errors, e.g. the collocation *wealthy repository* was not corrected to *rich repository* but to *wealth* (CO2). And a few more with register errors, e.g. the degree modifier *pretty*, which should have been corrected as *rather* or *somewhat*, was changed to *highly* or *very* (RE8). The inadequate changes that were grammatically incorrect mostly concerned prepositional errors, e.g. the preposition *to* after the verb *integrate*, which should have been corrected to *into*, was changed to *across* or *for* (PR4). Some also occurred with grammatical errors, e.g. instead of replacing the gerund *at achieving* after the verb *enable* with the *to*-infinitive *to achieve*, only the preposition *at* was changed to *in* (GR9). Only one inadequate change occurred with a register error: the combination "make … better" should have been corrected to *improved* but was instead changed to "make … improve" (RE10).

Easily confusable words were rarely deleted, especially not entirely. Instances in which changes for these words were coded as error deletions occurred because of extensive rephrasing.

There are two error deletions from the field of prepositions that commonly occurred with many tools. The first one is the erroneous phrase "the majority from" (PR3), which was often changed to *most*, thereby deleting the erroneous preposition

from. This deletion was judged as adequate as it did not alter the meaning but simply served to shorten the phrase. The second one is the erroneous combination "underpinnings from" (PR10), in which the noun *underpinnings* was often changed to *foundations*. Since the nouns have a similar meaning, these instances were also judged as adequate as the deletion merely changed the wording.

Two error deletions from the field of collocations and one from the field of grammar commonly occurred with many tools, namely the collocations "surfacing technologies" (CO6) and "deep implications" (CO7) as well as the phrase "is indicative enough" (GR10). Instead of replacing *surfacing* with *emerging* and *deep* with *profound* and altering the word order to "is indicative enough", some tools deleted the adjectives *surfacing* and *deep* and the adverb *enough*. These deletions were judged as inadequate as they led to a loss of information.

Two error deletions from the field of register commonly occurred with many tools. The first one is the phrase "pretty innovative" (RE8), in which the tools either deleted the degree modifier *pretty* entirely or replaced the phrase with *novel* or *unique*. Both changes were judged as inadequate as the first one led to a loss of information and the second one changed the original meaning. The second one is the personal pronoun *your* (RE9), which should have been corrected to *one's*, but was instead deleted by some tools. This change was considered inadequate as it led to a loss of information.

5 Discussion

Overall, the goal of this study was to provide an exploratory evaluation to answer three research questions. With regard to RQ1, the results indicate that, for the errors and texts at hand, there are clear differences between the overall correction performances of the AI tools investigated. Regarding *detection accuracy*, which determines how accurately the tools were able to detect errors in the texts, *GPT5* had the highest $F_{0.5}$ score. In terms of *correction suitability*, which reveals how suitable the tools' corrections actually were, *DeepL Write (default)* achieved the highest performance. However, this performance must be relativised. While the tool does find and adequately correct many errors, it also tends to rephrase entire chunks of sentences and, thereby, in some instances alters the original meaning or causes a loss of information. It also seems to make an inadequate or unnecessary change rather than leaving an error undetected, which is shown through the fact that the *detection accuracy* of its corrections was below the .70 threshold due to many additional and unnecessary changes made in the text. *DeepL Write (default)* might, therefore, be useful to eradicate errors but users should be aware that not all changes that the tool suggests are necessary. While *DeepL Write (default)* is the tool that rephrases the most during the correction process, other tools also sometimes alter more of the texts than strictly necessary. All in all, based on the analysis at hand, most tools appear to be more than capable of correcting errors; however, one must be careful and not blindly accept every change they make to a text.

In terms of RQ2 and contrary to Buck's (2025) assumption regarding the superiority of dedicated correction tools, in this particular analysis *GPT5*, one version of the multi-purpose AI tool, outperformed all specialised writing tools regarding the *detection accuracy* of their corrections. In agreement with Buck's (2025) assumption, both *GPT*s did not perform best but were surpassed by the dedicated writing tool *DeepL Write (default)* for the *correction suitability* of their corrections. Contrary to her suggestion, both *GPT*s did perform better than the other dedicated tools.

Regarding RQ3, for the specific error sample at hand the results reveal that there do seem to be differences between problem areas when it comes to the *correction suitability* of AI corrections, which fits in well with previous research (Fang et al. 2023; Lechien et al. 2023). Based on the results of this study, prepositional and grammatical errors seem to be rather easy for AI tools to detect and correct. Collocational errors are more difficult to rectify for all tools except for *DeepL Write (default)* and *GPT5* and only some tools perform well at spotting and correcting easily confusable words that have been used incorrectly. Register consistency in a text seems to be easiest to maintain for *DeepL Write (default)* and *Paperpal*. Interestingly, most inadequate changes that were made belong to the problem area confusable words, closely followed by register errors and collocational errors.

It is difficult to provide reasons for why specific errors were corrected or not. Some errors were corrected adequately in one text but not in a different one, even by the same tool. Therefore, while possible reasons will be discussed, it must be kept in mind that these are merely speculations based on this particular experiment that require further analysis and evaluation.

One reason might be the probabilistic and inconsistent nature of AI tools (Pfau et al. 2023; Y. Xu et al. 2024). While they can be used for text correction, the changes they make are not based on knowledge but on predicting which word is most likely to occur in which place. They lack understanding and semantic knowledge of a text, which might explain why confusable words, such as the false friends *concurrence* and *Konkurrenz* (CW9), are so hard to correct adequately and are prone to be corrected inadequately most commonly as well. Words that can occur in a certain context but do not carry the meaning the author wanted to convey will probably be corrected in the intended way only some of the time. This explanation, however, does not account for why collocational errors are so difficult for most AI tools to spot. After all, the aforementioned definition of collocations describes them as words that co-occur and "whose frequency of occurrence is larger than expected on the basis of chance" (Gries 2008: 6), which is similar to what LLMs are based on: predicting which word is most likely to occur after another.

Another possible reason for AI tools failing to find certain errors might be that performance rates decrease with increasing sentence difficulty and length (Fang et al. 2023; Wu et al., 2023; Y. Xu et al. 2024). In some cases when multiple errors were connected in the texts (e.g. T1: "ultimately ~~facilitating~~ enabling/helping

English language learners ~~at achieving~~ to achieve greater proficiency and confidence") error correction performance was worse. However, this speculation cannot be generalised based on the present study because the effect was not always found and not for all tools.

Since this kind of research is only in its infancy, this paper should be seen as a first exploratory analysis. Therefore, this paper does not claim to provide definite answers to the question of whether AI tools can be used as *artificial proofreaders* for academic texts. The results of this analysis should be evaluated cautiously as certain limitations must be kept in mind.

6 Limitations and future research

After discussing the results, this research must be put into perspective: Firstly, since AI tools are developed and updated rapidly and regularly, a replication of this study merely a few months post publication might yield different results. This is why this study does not provide an in-depth analysis of the correction performance for any specific errors. The main goal is to provide a general overview of the overall and problem-area-specific correction performance. Secondly, this paper will not go into detail regarding any ethical or data protection concerns that are involved in the use of AI for text correction purposes. There are very informative overviews on this topic (e.g. Bowen & Watson 2024; Buck 2025; Kohnke 2026; Son & Philpott 2026) that can be recommended to support an informed choice for or against as well as between different AI tools. And thirdly, many of the tools investigated in this study do not only offer text correction but also other features such as translation, rephrasing or text generation. This paper does not claim to evaluate the tools regarding their suitability for any other task.

Furthermore, this research has certain limitations that need to be kept in mind. As it was not the purpose of this study, an in-depth literature review in the form of a scoping review or meta-analysis of previous research on using AI tools for error correction was not carried out. Rather, a short and concise summary of what the author deemed to be some of the most relevant previous research for this paper was provided. Future research could aim to give such an in-depth overview to strengthen the theoretical foundation of this kind of research.

The generation of fictitious learner texts, rather than the use of authentic learner texts, can also be called into question. As this particular paper was planned as an exploratory foray into the field with the aim of finding out whether this kind of research would even yield interesting results, complete authenticity was – for the time – passed over in favour of a more controlled environment: having the same errors in different positions in different texts was judged to be an appropriate approach for this analysis. Especially because authentic learner texts are rife with expressions that are in a grey zone and cannot clearly be judged as errors. Nevertheless, the texts still aimed to resemble authentic learner language, and the errors were chosen because they could have been produced by German learners of

English. Since this controlled experiment has shown that there do seem to be differences between AI tools as well as error types regarding *detection accuracy* and *correction suitability*, further research should expand this methodology to authentic learner texts.

The errors were chosen based on previous literature (e.g. Alexander 1994; Fitikides 2002; Parkes 2003) as well as the author's own experience with texts written by German learners of English and were validated as actual errors using a corpus linguistic approach. As explained above, only a limited number of errors was used to facilitate a controlled environment for this exploratory study. Future research could aim to validate the errors as authentic by also using learner corpus data or previous empirical studies. Furthermore, in the controlled environment of the present experiment, all 50 errors were assumed to be equally salient across the three learner texts. In authentic texts, there are unpredictable error clusters and errors might interact with each other. Future research should investigate such different distributions in authentic learner texts and how well AI tools perform under these circumstances. Additionally, the present study does not differentiate between the severity of the errors. For learners, some errors are more difficult than others. Using learner judgements regarding the severity of different errors, future research could investigate whether AI tools might be especially helpful when it comes to more difficult errors. As it uses only a limited number of rather specific errors, the register analysis of this study might be seen as a bit too narrow. There are, of course, many more register errors that could be analysed. The examples in the present study were chosen because they could easily be classified as such using a corpus analysis. A follow-up study might broaden the scope of register errors, maybe even based on prescriptive rules, and investigate whether AI tools are able to adhere to those. Finally, for the rather controlled environment of this experiment descriptive statistics were deemed appropriate. Future research, using a larger and more representative sample, might include inferential statistics as well as try and find causalities and make broader generalisations.

The focus on German learner errors was deliberate but raises questions regarding the generalisation of the results to other ESL, EAL and EFL users. While the errors themselves might differ as learners with a different language background will presumably exhibit different L1 interference patterns, the error types will be similar. Other learners of English will also struggle with using prepositions and collocations correctly, distinguishing between confusable words, applying grammar rules correctly and adhering to a certain register. Therefore, the results from the present study can also be of interest to learners from different language backgrounds, but future research will need to conduct similar analyses for other languages in order to find out how well AI tools work for speakers with other L1s.

7 Conclusion

The findings of the present study highlight the differences between AI tools regarding their overall and problem-area-specific correction rates. These might be due to different LLMs that the tools are based on, differences in training, correction preferences for specific problem areas, etc. The tools with the highest overall performance rates can be seen as advantageous, especially for learners of English struggling with writing error-free texts. What can be also seen is that no tool achieves a perfect score and that human intervention and evaluation of the output is still necessary (Pfau et al. 2023; Y. Xu et al. 2024). Nevertheless, these tools can be useful to first and foremost detect errors in one's own text. This research has also shown how linguistic corpora can be used to evaluate certain words, phrases or structures regarding their appropriateness in a specific context[20].

This research can be seen as an experimental foray into this – as of yet – still rather new and under-researched field. Further research is not only recommended but required as more studies are necessary and vital to support, verify or refute the findings and, thereby, to provide more elaborate answers as to why these tools perform the way they do. One field of such future research might investigate one of the problem areas in more detail and find reasons for why errors from this problem area are easy to fix or not and whether the tendencies found in this analysis also apply for other errors. Another field of future research might be a diachronic analysis of the correction performance of different tools. By using the same fictitious learner texts and having the same and maybe additional tools correct them a few months or years from now, conclusions could be drawn regarding possible improvements implemented in the tools. Additional changes made by the AI tools were briefly mentioned. Future research might also investigate these surplus changes more closely. Can patterns be found? Do some tools have a preference for specific formulations and rigorously adapt texts accordingly regardless of whether the original phrase was just as acceptable?

This paper has shown that while the tools have potential as *artificial proofreaders*, not all corrections should be accepted without question. Therefore, a critical evaluation of AI applications as one aspect of *AI literacy* (D. T. K. Ng et al. 2021; OECD 2025) is a necessary future skill that both students and researchers need to acquire in order to responsibly and competently use these emerging technologies. This paper provides a basis for further research in this area and offers

[20] Acquiring *corpus literacy* and learning how to use corpora for such practical applications is not only recommended for linguists but can be advantageous for both students and teachers of English (e.g. Mair 2002; Timmis 2015) as well as anyone interested in the English language. *Corpus literacy* can, for example, be achieved by completing the freely available KorPLUS self-learning package. The development of this online course as well as the research this paper is based on was supported by the *Stiftung Innovation in der Hochschullehre* (Project *Digitale Kulturen der Lehre entwickeln – DiKuLe*, University of Bamberg).

practical insights for (non-native) users of English who wish to leverage AI tools effectively in academic writing.

Funding

The projects in which the present study has been realised (*Digitale Kulturen der Lehre entwickeln – DiKuLe*, University of Bamberg, and *Bamberger Kulturen der Lehre gemeinsam gestalten – BaKuLe*, University of Bamberg) are funded by the *Stiftung Innovation in der Hochschullehre*.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this contribution.

Acknowledgments

The author acknowledges the use of different dedicated AI tools and a multi-purpose AI tool in order to evaluate their potential as correction aids as specified in 3 Methods.

The author wishes to acknowledge the feedback and technological support from colleagues and a student at the University of Bamberg, in particular Prof. Dr. Julia Schlüter, Dr. Lukas Sönning, Claudia Schnellbögl, Lea Bracke, and Christopher Groß. The author would also like to thank the anonymous reviewers for their feedback and suggestions.

References

Alexander, Louis George. 1994. *Right word, wrong word: Words and structures confused and misused by learners of English*. Harlow: Longman Group UK Limited.

Amano, Tatsuya & Ramírez-Castañeda, Valeria & Berdejo-Espinola, Violeta & Borokini, Israel & Chowdhury, Shawan & Golivets, Marina & González-Trujillo, Juan David & Montaño-Centellas, Flavia & Paudel, Kumar & White, Rachel Louise et al. 2023. The manifold costs of being a non-native English speaker in science. *PLoS Biology* 21(7). https://doi.org/10.1371/journal.pbio.3002184

Ammon, Ulrich. 1996. The European Union (EU – formerly European Community): Status change of English during the last fifty years. In Fishman, Joshua A. & Conrad, Andrew W. & Rubal-Lopez, Alma (eds.), *Post-Imperial English*, 241–268. Berlin, New York: Mouton De Gruyter. https://doi.org/10.1515/9783110872187.241

Baumann, Andreas & Mühlenbernd, Roland. 2022. Less of the same: Modeling horror aequi and extravagance as mechanisms of negative frequency dependence in linguistic diversification. In Ravignani, Andrea & Asano, Rie & Valente, Daria & Ferretti, Francesco & Hartmann, Stefan & Hayashi, Misato & Jadoul, Yannick & Martins, Mauricio & Oseki, Yohei & Rodrigues, Evelina D. et al. (eds.), *The evolution of language: Proceedings of the joint conference on language evolution (JCoLE)*, 50–57. Nijmegen: Joint Conference on Language Evolution (JCoLE). https://doi.org/10.17617/2.3398549

Bo, Zhou & Pek, Lim Seong & Jian, Li & Cong, Wang & Nan, Lu Tian. 2025. The use of artificial intelligence tools in English academic writing among university students: A scoping review. *Language Teaching Research Quarterly* 53, 95–114. https://doi.org/10.32038/ltrq.2025.53.05

Bowen, José Antonia & Watson, C. Edward. 2024. *Teaching with AI: A practical guide to a new era of human learning*. Baltimore: Johns Hopkins University Press. https://doi.org/10.56021/9781421449227

Bridle, Marcus. 2019. Learner use of a corpus as a reference tool in error correction: Factors influencing consultation and success. *Journal of English for Academic Purposes* 37, 52-69. https://doi.org/10.1016/j.jeap.2018.11.003

Brown, Tom B. & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda et al. 2020. Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

Buck, Isabella. 2025. *Wissenschaftliches Schreiben mit KI*. Tübingen: UVK Verlag. https://doi.org/10.36198/9783838563657

Cengiz, Behice Ceyda & Bilki, Zeynep & Ataş, Amine Hatun & Celik, Berkan. 2025. Exploring second language writers' engagement with ChatGPT

feedback: Revision behaviors and perceptions. *System* 134. https://doi.org/10.1016/j.system.2025.103837

Clarke, Roger. 2023. The re-conception of AI: Beyond artificial, and beyond intelligence. *IEEE Transactions on Technology and Society* 4(1), 24–33. https://doi.org/10.1109/TTS.2023.3234051

Coyne, Steven & Sakaguchi, Keisuke & Galvan-Sosa, Diana & Zock, Michael & Inui, Kentaro. 2023. Analyzing the performance of GPT-3.5 and GPT-4 in grammatical error correction. *arXiv*. https://doi.org/10.48550/arXiv.2303.14342

Crossley, Scott A. & Bradfield, Franklin & Bustamente, Analynn. 2019. Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research* 11(2), 251–270. https://doi.org/10.17239/jowr-2019.11.02.01

Cummings, Robert E. & Monroe, Stephen M. & Watkins, Marc. 2024. Generative AI in first-year writing: An early analysis of affordances, limitations, and a framework for the future. *Computers and Composition* 71. https://doi.org/10.1016/j.compcom.2024.102827

Curry, Mary Jane & Lillis, Theresa. 2007. The dominance of English in global scholarly publishing. *International Higher Education* 46, 6–7. https://doi.org/10.6017/ihe.2007.46.7948

Davies, Mark. 2004. *British National Corpus (BNC) (from Oxford University Press)*. https://www.english-corpora.org/bnc/

Davies, Mark. 2008-. *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/

Davies, Mark. 2013. *Corpus of Global Web-Based English (GloWbE)*. https://www.english-corpora.org/glowbe/

Dolgova, Natalia & Mueller, Charles. 2019. How useful are corpus tools for error correction? Insights from learner data. *Journal of English for Academic Purposes* 39, 97–108. https://doi.org/10.1016/j.jeap.2019.03.007

Fang, Tao & Yang, Shu & Lan, Kaixin & Wong, Derek F. & Hu, Jinpeng & Chao, Lidia S. & Zhang, Yue. 2023. Is ChatGPT a highly fluent grammatical error correction system? A comprehensive evaluation. *arXiv*. https://doi.org/10.48550/arXiv.2304.01746

Fitikides, Timothy J. 2002. *Common mistakes in English with exercises*. 6th edn. Harlow: Pearson Education Limited.

Flowerdew, John. 1999a. Writing for scholarly publication in English: The case of Hong Kong. *Journal of Second Language Writing* 8(2), 123–145. https://doi.org/10.1016/S1060-3743(99)80125-8

Flowerdew, John. 1999b. Problems in writing for scholarly publication in English: The case of Hong Kong. *Journal of Second Language Writing* 8(3), 243–264. https://doi.org/10.1016/S1060-3743(99)80116-7

Flowerdew, John. 2008. Scholarly writers who use English as an additional language: What can Goffman's "stigma" tell us? *Journal of English for Academic Purposes* 7(2), 77–86. https://doi.org/10.1016/j.jeap.2008.03.002

Flowerdew, John. 2019. The linguistic disadvantage of scholars who write in English as an additional language: Myth or reality. *Language Teaching* 52(2), 249–260. https://doi.org/10.1017/S0261444819000041

Flowerdew, Lynne. 2022. Using corpora for writing instruction. In O'Keeffe, Anne & McCarthy, Michael J. (eds.), *The Routledge handbook of corpus linguistics*, 2nd edn., 443–455. London: Routledge. https://doi.org/10.4324/9780367076399-31

Gilmore, Audrey & Carson, David & Perry, Chad. 2006. Academic publishing: Best practice for editors, guest editors, authors and reviewers. *European Business Review* 18(6). 468–478. https://doi.org/10.1108/09555340610711094

Giray, Louie. 2023. Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering* 51, 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Gries, Stefan Th. 2008. Phraseology and linguistic theory: A brief survey. In Granger, Sylviane & Meunier, Fanny (eds.), *Phraseology: An interdisciplinary perspective*, 3–25. Amsterdam, Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/z.139.06gri

Hanauer, David I. & Englander, Karen. 2011. Quantifying the burden of writing research articles in a second language: Data from Mexican scientists. *Written Communication* 28(4). 403–416. https://doi.org/10.1177/0741088311420056

Hanauer, David I. & Sheridan, Cheryl L. & Englander, Karen. 2019. Linguistic injustice in the writing of research articles in English as a second language: Data from Taiwanese and Mexican researchers. *Written Communication* 36(1), 136–154. https://doi.org/10.1177/0741088318804821

Hüsch, Marc & Horstmann, Nina & Breiter, Andreas. 2025. *Künstliche Intelligenz im Studium – die Sicht von Studierenden im Wintersemester 2024/25*. CHE Hochschuldaten. https://hochschuldaten.che.de/kuenstliche-intelligenz-im-studium-die-sicht-von-studierenden-im-wintersemester-2024-25/ (last accessed on 22/12/2025).

Hwang, Haerim & Chang, Xingyu & Sun, Jiaxin. 2025. Generative AI is useful for second language writing, but when, why, and for how long do learners use it? *Journal of Second Language Writing* 69. https://doi.org/10.1016/j.jslw.2025.101230

Hyland, Ken. 2016a. 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

Hyland, Ken. 2016b. Language myths and publishing mysteries: A response to Politzer-Ahles et al. *Journal of Second Language Writing* 34, 9–11. https://doi.org/10.1016/j.jslw.2016.09.001

Imran, Muhammad & Almusharraf, Norah. 2023. Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology* 15(4). https://doi.org/10.30935/cedtech/13605

Kohnke, Lucas. *Microlearning with GenAI: Enhancing second language education*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009642859

Koltovskaia, Svetlana & Rahmati, Payam & Saeli, Hooman. 2024. Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement. *Journal of Second Language Writing* 65. https://doi.org/10.1016/j.jslw.2024.101130

Leacock, Claudia & Chodorow, Martin & Gamon, Michael & Tetreault, Joel. 2014. *Automated grammatical error detection for language learners*. 2nd edn. San Rafael: Morgan & Claypool. https://doi.org/10.1007/978-3-031-02153-4

Lechien, Jerome R. & Gorton, Amy & Robertson, Jean & Vaira, Luigi A. 2023. Is ChatGPT-4 accurate in proofread a manuscript in otolaryngology-head and neck surgery? *Otolaryngology-Head and Neck Surgery* 170(6), 1527–1530. https://doi.org/10.1002/ohn.526

Leitgeb, Thomas & Leitgeb, Michael. 2024. Empfehlungen für eine evidenzbasierte Nutzung von Künstlicher Intelligenz und Large Language Modells in Hochschulen: Ergebnisse eines systematischen Reviews. *phpublico – Fachzeitschrift für Bildung und Erziehung*, 1–34. https://doi.org/10.5281/ZENODO.13918915

Lennon, Paul. 1991. Error: Some problems of definition, identification, and distinction. *Applied Linguistics* 12(2), 180–196. https://doi.org/10.1093/applin/12.2.180

Li, Mingyu & Wang, Fengqiang. 2026. EAL scholars' experiences of writing for publication in English: A meta-ethnographic synthesis of qualitative evidence. *English for Specific Purposes* 81, 16–31. https://doi.org/10.1016/j.esp.2025.09.001

Liu, Dilin. 2013a. Salience and construal in the use of synonymy: A study of two sets of near-synonymous nouns. *Cognitive Linguistics* 24(1), 67–113. https://doi.org/10.1515/cog-2013-0003

Liu, Dilin. 2013b. Using corpora to help teach difficult-to-distinguish English words. *English Teaching* 68(3). 27–50. https://doi.org/10.15858/engtea.68.3.201309.27

Liu, Dilin & Lei, Lei. 2017. *Using corpora for language learning and teaching*. Alexandria, Virginia: tesol press.

Lo, Chung Kwan. 2023. What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences* 13(4). https://doi.org/10.3390/educsci13040410

López-Navarro, Irene & Moreno, Ana I. & Quintanilla, Miguel Ángel & Rey-Rocha, Jesús. 2015. Why do I publish research articles in English instead of

my own language? Differences in Spanish researchers' motivations across scientific domains. *Scientometrics* 103, 939–976. https://doi.org/10.1007/s11192-015-1570-1

Mair, Christian. 2002. Empowering non-native speakers: The hidden surplus value of corpora in continental English departments. In Kettemann, Bernhard & Marko, Georg (eds.), *Teaching and learning by doing corpus analysis: Proceedings of the fourth international conference on teaching and language corpora, Graz 19-24 July, 2000*, 119–130. Leiden: De Gruyter Brill. https://doi.org/10.1163/9789004334236_011

McKinley, Jim & Rose, Heath. 2018. Conceptualizations of language errors, standards, norms and nativeness in English for research publication purposes: An analysis of journal submission guidelines. *Journal of Second Language Writing* 42, 1–11. https://doi.org/10.1016/j.jslw.2018.07.003

Ng, Davy Tsz Kit & Leung, Jac Ka Leung & Chu, Samuel Kai Wah & Qiao, Maggie Shen. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2. https://doi.org/10.1016/j.caeai.2021.100041

Ng, Hwee Tou & Wu, Siew Mei & Briscoe, Ted & Hadiwinoto, Christian & Susanto, Raymond Hendy & Bryant, Christopher. 2014. The CoNLL-2014 shared task on grammatical error correction. In Ng, Hwee Tou & Wu, Siew Mei & Briscoe, Ted & Hadiwinoto, Christian & Susanto, Raymond Hendy & Bryant, Christopher (eds.), *Proceedings of the eighteenth conference on computational natural language learning: Shared task*, 1–14. Baltimore, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1701

OECD. 2025. *Empowering learners for the age of AI: An AI literacy framework for primary and secondary education (Review draft)*. OECD. https://ailiteracyframework.org/

Ouyang, Shuyin & Zhang, Jie M. & Harman, Mark & Wang, Meng. 2025. An empirical study of the non-determinism of ChatGPT in code generation. *ACM Transactions on Software Engineering and Methodology* 34(2), 1–28. https://doi.org/10.1145/3697010

Parkes, Geoff. 2003. *The mistakes clinic for German-speaking learners of English: Error correction exercises and detailed error analysis for German-speaking students and teachers of English* (Reprinted with corr.). Southampton: Englang Books.

Pfau, Adam & Polio, Charlene & Xu, Yiran. 2023. Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics* 2(3). https://doi.org/10.1016/j.rmal.2023.100083

Politzer-Ahles, Stephen & Girolamo, Teresa & Ghali, Samantha. 2020. Preliminary evidence of linguistic bias in academic reviewing. *Journal of English for Academic Purposes* 47. https://doi.org/10.1016/j.jeap.2020.100895

Quinn, Cynthia. 2015. Training L2 writers to reference corpora as a self-correction tool. *ELT Journal* 69(2), 165-177. https://doi.org/10.1093/elt/ccu062

Quirk, Randolph & Greenbaum, Sidney & Leech, Geoffrey & Svartvik, Jan. 1985. *A comprehensive grammar of the English language*. London: Longman.

Ramírez-Castañeda, Valeria. 2020. Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: The case of Colombian researchers in biological sciences. *PLoS ONE* 15(9). https://doi.org/10.1371/journal.pone.0238372

Rankin, Tom & Schiftner, Barbara. 2011. Marginal prepositions in learner English: Applying local corpus data. *International Journal of Corpus Linguistics* 16(3), 412–434. https://doi.org/10.1075/ijcl.16.3.07ran

Rich, Elaine & Knight, Kevin. 1991. *Artificial intelligence*. 2nd edn. New York: McGraw-Hill.

Rohdenburg, Günter. 2003. Cognitive complexity and horror aequi as factors determining the use of interrogative clause linkers in English. In Rohdenburg, Günter & Mondorf, Britta (eds.), *Determinants of grammatical variation in English*, 205–250. Berlin, New York: Mouton de Gruyter. https://doi.org/10.1515/9783110900019.205

Schlüter, Julia & Deckert, Katharina. 2025. Artificial Intelligence vs. Corpus Literacy: Ansätze zur Vermittlung reflektierter Schreibkompetenz in der Wissenschaftssprache Englisch. In Mrohs, Lorenz & Franz, Julia & Herrmann, Dominik & Lindner, Konstantin & Staake, Thorsten (eds.), *Digitales Lehren und Lernen an der Hochschule: Strategien – Bedingungen – Umsetzung*, 111–119. Bielefeld: transcript Verlag. https://doi.org/10.1515/9783839471203-011

Sheldon, Elena. 2020. 'We cannot abandon the two worlds, we have to be in both': Chilean scholars' views on publishing in English and Spanish. *Journal of English for Research Publication Purposes* 1(2), 120–142. https://doi.org/10.1075/jerpp.19016.she

Son, Jeong-Bae & Philpott, Andrew. 2026. *AI-Powered language teaching*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009650809

Staples, Shelley & Egbert, Jesse & Biber, Douglas & Conrad, Susan. 2015. Register variation: A corpus approach. In Tannen, Deborah & Hamilton, Heidi E. & Schiffrin, Deborah (eds.), *The Handbook of Discourse Analysis*, 2nd edn., 505–525. West Sussex: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118584194.ch24

Strauss, Pat. 2019. Shakespeare and the English poets: The influence of native speaking English reviewers on the acceptance of journal articles. *Publications* 7(1). https://doi.org/10.3390/publications7010020

Tan, Xiao. 2025. Academic socialization with generative AI: A longitudinal case study of an L2 graduate student's academic literacies development. *Journal of Second Language Writing* 70. https://doi.org/10.1016/j.jslw.2025.101263

Timmis, Ivor. 2015. Corpora in perspective. In *Corpus linguistics for ELT: research and practice*, 183–195. London, New York: Routledge, Taylor & Francis Group. https://doi.org/10.4324/9781315715537

Turner, Joan. 2004. Language as academic purpose. *Journal of English for Academic Purposes* 3(2), 95–109. https://doi.org/10.1016/S1475-1585(03)00054-7

Uzuner, Sedef. 2008. Multilingual scholars' participation in core/global academic communities: A literature review. *Journal of English for Academic Purposes* 7(4), 250–263. https://doi.org/10.1016/j.jeap.2008.10.007

Warschauer, Mark & Tseng, Waverly & Yim, Soobin & Webster, Thomas & Jacob, Sharin & Du, Qian & Tate, Tamara. 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. https://doi.org/10.1016/j.jslw.2023.101071

Wu, Haoran & Wang, Wenxuan & Wan, Yuxuan & Jiao, Wenxiang & Lyu, Michael. 2023. ChatGPT or Grammarly? Evaluating ChatGPT on grammatical error correction benchmark. *arXiv*. https://doi.org/10.48550/arXiv.2303.13648

Xu, Jian. 2025. Harnessing AI in academic writing: The complex interplay of AI literacy and self-directed learning among university L2 learners. *System* 135. https://doi.org/10.1016/j.system.2025.103855

Xu, Yiran & Polio, Charlene & Pfau, Adam. 2024. Optimizing AI for assessing L2 writing accuracy: An exploration of temperatures and prompts. In Chapelle, Carol A. & Beckett, Gulbahar H. & Ranalli, Jim (eds.), *Exploring Artificial Intelligence in Applied Linguistics*, 151–174. Iowa: Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.10

Investigated AI Tools

OpenAI. *GPT4o-mini*. Accessed in July 2024 (LT1), February 2025 (LT2, LT3) through https://hawki.its.uni-bamberg.de/.

OpenAI. *GPT5*. Accessed in August and September 2025 through chat.openai.com.

DeepL Write. Accessed in October 2025 through https://www.deepl.com/de/write.

Grammarly. Accessed in October 2025 through app.grammarly.com.

LanguageTool. Accessed in October 2025 through https://languagetool.org/de.

Mimir Mentor. Accessed in October 2025 through https://mimir-mentor.com/en/features-correction/.

Paperpal. Accessed in October 2025, through https://paperpal.com/.

ProWritingAid. Accessed in October 2025 through https://prowritingaid.com/grammar-checker.

Trinka.ai. Accessed in October 2025 through https://www.trinka.ai/grammar-checker.

Wortliga. Accessed in 2025 through https://wortliga.de/rechtschreibpruefung/.

Appendix

Error Glossary

Table 3: All errors included in the three learner texts along with their assigned problem areas, error codes, model corrections, and instances of error deletions and inadequate corrections.

	Error code	Error	Model correction(s)	Instances of error deletions		Instances of inadequate corrections
				adequate deletions	inadequate deletions	
Preposition usage	PR1	*has led into*	*has led to*	*has sparked*	*brought about*	
	PR2	*revolution at*	*revolution in*			
	PR3	*majority from*	*majority of*	*most*	*across/*phrase deleted	
	PR4	*INTEGRATE … to*	*INTEGRATE … into*			*INTEGRATE … across/for*
	PR5	*emergence from*	*emergence of*		sentence rephrased	
	PR6	*exemplified through*	*exemplified by*	*such as/as seen through/ demonstrated*		
	PR7	*approach at*	*approach to*		*a fresh perspective on*	*approach in*
	PR8	*potential from*	*potential of*	sentence rephrased		
	PR9	*through exploring*	*by exploring*	sentence rephrased	sentence rephrased	
	PR10	*underpinnings from*	*underpinnings of*	*foundations/fundamentals of/*sentence rephrased		
Collocati	CO1	*priceless insights*	*invaluable insights*			
	CO2	*wealthy repository*	*rich repository*			*wealth/valuable repository*
	CO3	*elementary part*	*fundamental part*			*elemental part/elementary aspect*
	CO4	*ENHANCE the importance*	*EMPHASIZE/STRESS/HIGHLIGHT/ UNDERSCORE the importance*			
	CO5	*running dialogue*	*ongoing dialogue*			*current dialogue*
	CO6	*surfacing technologies*	*emerging/developing technologies*		*technologies*	*emergent*
	CO7	*deep implications*	*profound/important/significant implications*		*implications*	*far-reaching*
	CO8	*FACILITATE … learners*	*ENABLE/HELP … learners*			

ons	CO9	*meaningful element*	*important element*	*integral/meaningful part/* sentence rephrased		*significant element*
	CO10	*insights … SUPPLY*	*insights … PROVIDE/OFFER/GIVE*	sentence rephrased	sentence rephrased	
Easily confusable words	CW1	*digitalized/digitalised*	*digital*			*digitized/digitised*
	CW2	*approximate*	*approach*	sentence rephrased		
	CW3	*procure*	*provide/give/offer*			*gain/obtain/produce*
	CW4	*INDIVIDUATE*	*INDIVIDUALIZE/INDIVIDUALISE*			*SUPPLEMENT/PERSONALIZE/AUGMENT/ COMPLEMENT/ENHANCE/TAILOR/IMPROVE*
	CW5	*intrigues*	*intricacies*		sentence rephrased	*challenges/nuances/mechanisms/issues/risks*
	CW6	*COMPLIMENT*	*COMPLEMENT*			*BUILD on*
	CW7	*significance*	*importance*			
	CW8	*LEVER*	*LEVERAGE*			
	CW9	*concurrence*	*competition/rivalry*		sentence rephrased	*convergence/adoption/combination/confluence/ advancement/integration/intersection/advent*
	CW10	*appliances*	*applications*	sentence rephrased	sentence rephrased	*use*
Grammaticalstru	GR1	*promise redefining*	*promise to redefine*			
	GR2	*the majority … is/provides*	*the majority … are/provide*			
	GR3	*are beginning integrating*	*are beginning to integrate*		*increasingly integrate/ are integrating*	
	GR4	*this paper is exploring*	*this paper explores*			
	GR5	*tendency of improving*	*tendency to improve*	*trend toward(s)*		*tendency toward(s) improving*
	GR6	*aims at shedding light*	*aims to shed light*			*aims to shedding*
	GR7	*discuss about*	*discuss*	sentence rephrased		

ctures	GR8	*the question how*	*the question <u>of</u> how*	*asks how/: How can*		
	GR9	*ENABLE … <u>at achieving</u>*	*ENABLE … <u>to achieve</u>*			*in achieving*
	GR10	*is <u>enough indicative</u>*	*is <u>indicative enough</u>*		*is indicative*	*is sufficiently/highly indicative*
Academic register	RE1	*a lot of*	*many/a host of*		deleted	*traditional*
	RE2	*<u>gotten</u> considered/<u>get</u> overlooked*	*<u>been</u> considered/<u>be</u> overlooked*	*are considered*		*cot considered*
	RE3	sentence-initial *But*	sentence-initial *However/Nevertheless/Nonetheless*		deleted	
	RE4	*<u>getting</u> adopted/improved*	*<u>being</u> adopted/improved*	*are adopting/improving*		*becoming/can be/may be/are*
	RE5	*there<u>'s</u>*	*there <u>is</u>*	rephrased		
	RE6	*gigantic*	*enormous/vast*		deleted	*significant/great/tremendous/huge/considerable*
	RE7	*shouldn<u>'t</u>*	*should <u>not</u>*	rephrased		*is not/cannot/must not*
	RE8	*<u>pretty</u> innovative*	*<u>rather/quite/somewhat/fairly</u> innovative*		*innovative/novel/unique*	*highly/truly/more/new/very/genuinely*
	RE9	*your*	*one's*	rephrased	deleted	*their/learners'/a patient's/English/students'/ones/a persons/an individual's/individual/patients/ patients'/academic*
	RE10	*MAKE … better*	*IMPROVE*			*MAKE … improve*

Text 1 (T1)

The ~~digitalized/digitalised~~ → digital [CW1] age has led ~~into~~ → to [PR1] a revolution ~~at~~ → in [PR2] language learning and literacy, introducing ~~a lot of~~ → many/a host of [RE1] innovative tools that promise ~~redefining~~ → to redefine [GR1] the way we ~~approximate~~ → approach [CW2] language learning. The majority ~~from~~ → of [PR3] language programs ~~is~~ → are [GR2] beginning ~~integrating~~ → to integrate [GR3] digital tools ~~to~~ → into [PR4] their curriculum, recognizing/recognising the ~~priceless~~ → invaluable [CO1] insights they ~~procure~~ → provide/give/offer [CW3] for personalized/personalised learning experiences. Among these tools, linguistic corpora have long ~~gotten~~ → been [RE2] considered a touchstone for researchers, teachers(,) and learners, providing a ~~wealthy~~ → rich [CO2] repository of examples of language use across genres, styles(,) and contexts. ~~But~~ → However/Nevertheless/Nonetheless [RE3], as we advance further into the 21st century, the emergence ~~from~~ → of [PR5] Artificial Intelligence (AI) technologies, exemplified ~~through~~ → by [PR6] ChatGPT, DeepL Write(,) and Grammarly, presents a new approach ~~at~~ → to [PR7] language learning and writing skill development. In response, these Large Language Models (LLMs) are increasingly ~~getting~~ → being [RE4] adopted by educators to ~~individuate~~ → individualize/individualise [CW4] traditional instruction, making AI literacy ~~an elementary~~ → a fundamental [CO3] part of language learning. This paper ~~is exploring~~ → explores [GR4] the dynamic interplay between corpora, which are well-established tools in linguistics, and the emerging potential ~~from~~ → of [PR8] AI-driven tools such as ChatGPT. ~~There's~~ → There is [RE5] a clear tendency for these tools ~~of improving~~ → to improve [GR5] English writing skills, especially for non-native English speakers.

For a university student, studying the ~~intrigues~~ → intricacies [CW5] of the English language and exploring how AI can ~~compliment~~ → complement [CW6] traditional corpus-based approaches is timely and relevant. As a native speaker of German, the transition to English proficiency presents ~~gigantic~~ → enormous/vast [RE6] challenges and opportunities that ~~enhance~~ → emphasize/stress/highlight/underscore [CO4] the ~~significance~~ → importance [CW7] of using corpora and AI. This paper argues that the advent of AI technologies ~~shouldn't~~ → should not [RE7] be seen as a replacement for corpus-based learning, but rather as an opportunity to ~~lever~~ → leverage [CW8] the strengths of both. ~~Through~~ → By [PR9] exploring the possibilities, limitations(,) and potential synergies between corpora and AI technologies, this study aims ~~at shedding~~ → to shed [GR6] light on ~~pretty~~ → rather/quite/somewhat/fairly [RE8] innovative strategies for improving ~~your~~ → one's [RE9] writing skills. Doing so will contribute to the ~~running~~ → ongoing [CO5] dialogue in applied linguistics and language education.

In navigating the ~~concurrence~~ → competition/rivalry [CW9] of traditional and ~~surfacing~~ → emerging/developing [CO6] technologies in language learning, this paper ~~discusses about~~ → discusses [GR7] the theoretical underpinnings ~~from~~ → of [PR10] corpus linguistics and AI, the practical ~~appliances~~ → applications [CW10] of these tools in language learning, and their ~~deep~~ → profound/important/significant [CO7] implications for learners, educators(,) and curriculum developers. This paper raises the question _ → of [GR8] how the fusion of corpus analysis and AI-driven feedback can create a more holistic and practical approach to ~~making~~ → improving [RE10] writing skills ~~better~~, ultimately ~~facilitating~~ → enabling/helping [CO8] English language learners ~~at achieving~~ → to achieve [GR9] greater proficiency and confidence. This particular development is ~~enough indicative~~ → indicative enough [GR10] of a broader trend in which data is becoming a ~~meaningful~~ → important [CO9] element of educational strategies, with teachers and students benefitting from the insights it ~~supplies~~ → provides/offers/gives [CO10].

Figure 2: Text 1 (T1) with errors, corrections and problem areas (original text generated with *GPT4o-mini*, revised to better fit the purpose).

Text 2 (T2)

In recent years, the digital revolution ~~at~~ → in [PR2] medicine has led ~~into~~ → to [PR1] many innovative advances that promise ~~redefining~~ → to redefine [GR1] our healthcare approach. Emerging technologies, exemplified ~~through~~ → by [PR6] digital health tools and data analytics, are ~~getting~~ → being [RE4] adopted by many healthcare institutions. This shift underscores the ~~significance~~ → importance [CW7] of integrating these technologies into ~~a lot of~~ → many/a host of [RE1] traditional medical practices to ~~procure~~ → provide/give/offer [CW3] insights into patient care.

The emergence ~~from~~ → of [PR5] digital tools has ~~gotten~~ → been [RE2] considered a significant milestone because they offer a ~~wealthy~~ → rich [CO2] repository of data that can be ~~levered~~ → leveraged [CW8] to better understand the ~~intrigues~~ → intricacies [CW5] of various diseases. However, the ~~concurrence~~ → competition/rivalry [CW9] of these technologies also brings with it the need to ~~discuss about~~ → discuss [GR7] the potential of these tools to ~~make~~ → improve [RE10] healthcare outcomes ~~better~~ in the future. This paper aims ~~at shedding~~ → to shed [GR6] light on how they can ~~individuate~~ → individualize/individualise [CW4] patient treatment plans and offer a ~~pretty~~ → rather/quite/somewhat/fairly [RE8] innovative approach to medical practice.

One of the ~~priceless~~ → invaluable [CO1] insights of the ~~digitalized/digitalised~~ → digital [CW1] revolution is in the area of personalized/personalised medicine. There is a clear tendency ~~of improving~~ → to improve [GR5] patient outcomes by tailoring treatments based on ~~your~~ → one's [RE9] genetic makeup, lifestyle, and other factors. ~~Through~~ → By [PR9] exploring the applications of these tools, this personalized/personalised approach ~~at~~ → to [PR7] medicine has shown ~~gigantic~~ → enormous/vast [RE6] promise in treating chronic diseases and enhancing the effectiveness of therapies. ~~But~~ → However/Nevertheless/Nonetheless [RE3] how to effectively integrate these digital tools ~~to~~ → into [PR4] the majority ~~from~~ → of [PR3] healthcare practices remains a fundamental part of the ongoing dialogue among medical professionals.

The promise to redefine medical education is also evident as educators are beginning ~~integrating~~ → to integrate [GR3] new technologies more fully into their students' learning environments. The majority of these new tools ~~provides~~ → provide [GR2] valuable insights to students and ~~facilitate~~ → enable/help [CO8] learners to receive a more personalized/personalised education. The importance of these findings ~~shoulda't~~ → should not [RE7] be underestimated, as they play a crucial role in shaping the future of medical education and enable eager learners ~~at achieving~~ → to achieve [GR9] a deeper understanding of the material. By ~~complimenting~~ → complementing [CW6] traditional teaching methods, digital tools can provide a more complete learning experience.

This paper ~~is exploring~~ → explores [GR4] the potential of digital technologies to revolutionize/revolutionise the medical field. It aims to shed light on the underpinnings ~~from~~ → of [PR10] this transformation and discuss the ~~deep~~ → profound/important/significant [CO7] implications for current and future medical practices. The approach of using these ~~surfacing~~ → emerging/developing [CO6] technologies to achieve better patient outcomes is ~~enough indicative~~ → indicative enough [GR10] of the impact they can have in the medical field. As such, the paper ~~enhances~~ → emphasizes/stresses/highlights/underscores [CO4] the importance of understanding the ~~appliances~~ → applications [CW10] of these tools and the innovations they bring to healthcare. ~~There's~~ → There is [RE5] quite a lot of evidence to suggest that new digital tools will become ~~an elementary~~ → a fundamental [CO3] part of the medical field.

In conclusion, the ~~running~~ → ongoing [CO5] dialogue around the adoption of digital technologies in medicine highlights the need to address the implications they bring as well as answer the question – → of [GR8] how they can be used effectively. The insights ~~supplied~~ → provided/offered/given [CO10] by these tools promise to redefine the way we ~~approximate~~ → approach [CW2] healthcare and make them a ~~meaningful~~ → important [CO9] element of medical practice. Through this exploration, we can better appreciate the potential ~~from~~ → of [PR8] digital innovation to improve patient care and medical education.

Figure 3: Text 2 (T2) with errors, corrections and problem areas (original text generated with *GPT4o-mini*, revised to better fit the purpose).

Text 3 (T3)

In recent years, there has been a so-called digital revolution ~~at~~ → **in** [PR2] various fields, with the legal sector being no exception. The emergence ~~from~~ → **of** [PR5] new technologies has led ~~into~~ → **to** [PR1] a profound shift in how legal practitioners ~~approximate~~ → **approach** [CW2] their work. ~~A lot of~~ → **Many/A host of** [RE1] scholars stress the ~~significance~~ → **importance** [CW7] of understanding these changes as they promise ~~redefining~~ → **to redefine** [GR1] the landscape of legal practice. ~~There's~~ → **There is** [RE5] an intricate relationship between ~~surfacing~~ → **emerging/developing** [CO6] technologies and the legal profession. This paper aims ~~at shedding~~ → **to shed** [GR6] light on how the majority ~~from~~ → **of** [PR3] these tools can enhance efficiency and ~~make~~ → **improve** [RE10] client service ~~better~~.

Due to the ~~running~~ → **ongoing** [CO5] dialogue between academia and practice, there is a growing tendency ~~of improving~~ → **to improve** [GR5] existing legal frameworks through the integration of digital solutions. Many law firms are beginning ~~integrating~~ → **to integrate** [GR3] artificial intelligence and machine learning into their operations, recognizing the potential of these innovations to provide ~~priceless~~ → **invaluable** [CO1] insights and streamline processes adjusted to ~~your~~ → **one's** [RE9] own needs. ~~But~~ → **However/Nevertheless/Nonetheless** [RE3] it has to be kept in mind that the underpinnings ~~from~~ → **of** [PR10] these technologies need further examination to fully grasp their ~~deep~~ → **profound/important/significant** [CO7] implications for the legal profession.

The ~~concurrence~~ → **competition/rivalry** [CW9] of traditional methods and these digital innovations, exemplified ~~through~~ → **by** [PR6] their applications, reveals a ~~wealthy~~ → **rich** [CO2] repository of both opportunities and challenges. To ~~lever~~ → **leverage** [CW8] the advantages of these tools effectively, legal professionals must adopt a ~~pretty~~ → **rather/quite/somewhat/fairly** [RE8] innovative approach ~~at~~ → **to** [PR7] their usage that not only enhances operational efficiency, but also ~~individuates~~ → **individualizes/individualises** [CW4] the client experience. ~~Through~~ → **By** [PR9] exploring these dynamics, this paper aims to highlight the importance of conducting research that can ~~procure~~ → **provide/give/offer** [CW3] more detailed insights into these tools and, consequently, of responsibly and sensibly integrating digital technologies ~~to~~ → **into** [PR4] legal education in order to ~~facilitate~~ → **enable/help** [CO8] learners to work more productively.

Furthermore, a ~~meaningful~~ → **important** [CO9] element to ~~discuss about~~ → **discuss** [GR7] is the possibilities that emerge from legal services ~~getting~~ → **being** [RE4] improved through technology. Insights ~~supplied~~ → **provided/offered/given** [CO10] by recent studies illustrate how the ~~appliances~~ → **applications** [CW10] of digital tools can enable learners ~~at achieving~~ → **to achieve** [GR9] a deeper understanding of the law that can ~~compliment~~ → **complement** [CW6] their already existing knowledge. As the aim of this paper ~~shouldn't~~ → **should not** [RE7] merely be to highlight the positive aspects, it is also imperative to discuss the ~~intrigues~~ → **intricacies** [CW5] involved and the ethical considerations that should not ~~get~~ → **be** [RE2] overlooked.

Ultimately, the potential ~~from~~ → **of** [PR8] emerging technologies in the legal field is ~~gigantic~~ → **enormous/vast** [RE6] and they will probably be ~~an elementary~~ → **a fundamental** [CO3] part of education going forward. The question _ → **of** [GR8] how to balance innovation with the fundamental principles of law is ~~enough indicative~~ → **indicative enough** [GR10] of the challenges ahead. As this paper ~~is exploring~~ → **explores** [GR4] these themes, it will ~~enhance~~ → **emphasize/stress/highlight/underscore** [CO4] the importance of adapting legal education and practice to the changing landscape to ensure that at least the majority of legal professionals ~~is~~ → **are** [GR2] equipped with the knowledge and skills necessary to thrive in this ~~digitalized/digitalised~~ → **digital** [CW1] age.

Figure 4: Text 3 (T3) with errors, corrections and problem areas (original text generated with *GPT4o-mini*, revised to better fit the purpose).

About the journal

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses is a Diamond Open Access journal supported by PUBLIA – SLUB Open Publishing.

DOI https://doi.org/10.62408/ai-ling

ISSN 2943-0070