"AI, will you help?" How learners use Artificial Intelligence when writing

Jarosław Krajka – Izabela Olszak

DOI: 10.18355/XL.2024.17.01.03

Abstract
The current surge in the exploitation of Artificial Intelligence (AI) tools, accompanied by the subsequent rise in popularity of AI across different spheres of life, has unlocked a multitude of opportunities for students to employ diverse AI strategies to augment their process of learning. Drawing on the proliferation of Artificial Intelligence, the article examines the process of assisting academic writing instruction with an AI-enhanced word processor. The purpose of the paper is to investigate how successful advanced students are in determining whether an essay was written by AI tools or by a human and how much AI assistance they need to summarize, generate text and write from prompts when trained to use an AI-assisted word processor. The empirical data for the scientific investigation was obtained through a quasi-experimental treatment involving a single group of undergraduate applied linguistics students. The findings of this study indicate the high linguistic sensitivity of the research participants to factors regarding language and layout, which allows them to distinguish human authors from AI-powered texts. The current investigation possesses potential advantages for educators in the realm of foreign language acquisition and instruction, as they contemplate the strengths of their bilingual language learners within academic writing instruction.
Key words: generative Artificial Intelligence; burstiness; perplexity; English for Academic purposes; writing instruction

Introduction

Recently, the third technological revolution of Artificial Intelligence, following the development of the Internet and personal miniaturized computers, has brought about new developments, improvements, or features used for numerous purposes including reading, writing, artistic development as well as entertainment. A new era for teaching and learning in all fields of language instruction has begun with the appearance and widespread use of Artificial Intelligence technologies, best epitomized by the well-known Chat-GPT text interaction tool.

Previous research has shown that EFL learners often do not have sufficient latitude to output at a satisfactory level when writing in a second language. Technology and AI-enhanced tools open up a myriad of opportunities for language acquisition. Digitalization of the learning process extends beyond technical assistance and begins to mimic face-to-face language instruction. Such virtual engagement broadens the scope of communication and adds appeal to language and cultural immersion.

The purpose of the current study was to examine to what extent advanced students make use of the opportunities to summarize, generate text, and write from prompts when trained to use an AI-assisted word processor. At the same time, the focus of the study was to see how successful advanced learners of English can be at distinguishing human from AI-generated writing, based on the concepts of burstiness and perplexity.

Literature review

Artificial Intelligence (AI) and Large Language Models (LLMs) in the writing classroom

With the concept of Artificial Intelligence becoming more and more prevalent in all areas of life, it is no wonder that currently available AI tools quickly find their way to the language classroom. Indeed, potential benefits and implementation possibilities need to be considered concerning the definitions and characteristics of these tools. Classical definitions view AI "as computer systems that have been designed to interact with the world through capabilities (for example, visual perception and speech recognition) and intelligent behaviours (for example, assessing the available information and then taking the most sensible action to achieve a stated goal) that we would think of as essentially human" (Luckin et al., 2016: 14). These computer systems include a wide range of technologies and methods, such as machine learning, adaptive learning, natural language processing, data mining, crowdsourcing, neural networks or algorithms (Pokrivcakova, 2019). AI-powered tools are applied in computer linguistics, in the creation of computer languages, machine translations and improvement of human-machine communication via speech recognition and speech synthesis. As they also belong to the currently emerging fields in educational technology, many authors see significant benefits they could bring both to students and teachers.

As Baker and Smith (2019) state, AI tools used in education can be of three different kinds:

a. Learner-facing AI tools are software that students employ to learn a subject matter.
b. Teacher-facing systems are utilised by teachers to reduce workload and make their output more effective in specific tasks, such as administration, assessment, feedback and plagiarism detection.
c. System-facing AI tools provide information for administrators and managers on the institutional level, for example, they help monitor attrition patterns across faculties or colleges.

The current study focuses only on the first group, on those applications and services that are within the control of the learner and that form a personal learning environment.

Artificial Intelligence tools operate based on language modelling (LM), which, according to Zhao et al. (2023), is one of the major approaches to advancing language intelligence of machines which aims to model the generative likelihood of word sequences to predict the probabilities of future or missing words. As characterized by Sejnowski (2023), Large Language Models are transformative, pre-trained foundational models that are self-supervised and can be adapted with fine-tuning to a wide range of natural language tasks, each of which would have previously required a separate network model. For instance, a widely acclaimed GPT-3 can carry on dialogues with humans on many topics after minimal priming (feeding) with a few examples. To be more specific, the process of priming enables an AI tool to teach itself to "speak" English by "reading" text (Sejnowski, 2023).

Even though AI use in language education is still in its infancy, some authors have already provided comprehensive descriptions of its application for various purposes, including writing skills. Pokrivcakova (2019) sees AI applications as tools for generating personalized learning materials, writing assistants, and conversational partners. Especially when enhanced with an avatar "body" through virtual reality, all of these can create authentic virtual reality and game-based learning

environments. Virtual agents can act as teachers, facilitators or students' peers, engaging in meaningful conversations with learners as Intelligent Personal Assistants (Frazier et al., 2020). Such intelligent conversational partners are important assistance or even replacement for the teacher in the process of writing, especially when a student writer becomes stuck, is not sure of how to use certain forms or fails to understand a certain passage or task.

A wide range of uses of AI tools and LLMs in writing instruction is given, among others, by Kasneci et al. (2023), who classify possible applications taking the age of students into account. As regards advanced learners at the tertiary level, large language models can assist in the research and writing tasks, development of critical thinking and problem-solving skills, generation of summaries and outlines of text, hinting at unexplored aspects and current research topics and helping to better understand and analyze the material. Of particular interest is the capacity of such tools for analyzing student's writing and providing tailored feedback on the writing input. Kasneci et al. (2023) as well as Bonner et al. (2023) also advocate using LLMs with advanced learners in academic writing for identifying and correcting typos, highlighting (potential) grammatical inconsistencies, generating summaries and outlines of challenging texts, identifying possibilities for topic-specific style improvement and suggesting adequate improvement strategies in research essays.

Perplexity and burstiness in written texts

When distinguishing human from AI-generated text, the notions of burstiness and perplexity help understand the patterns and complexities in the text. The former refers to the frequency of rare words or phrases appearing in a text, in other words, indicates how often uncommon words are used. A text with high burstiness has a greater number of rare words or phrases, while a text with low burstiness uses more common words and phrases (Alexander, 2023). As Kuribayashi et al. (2021) claim, burstiness measures how predictable a piece of content is by the homogeneity of the length and structure of sentences throughout the text. This means that burstiness is parallel to perplexity, however, at the level of sentences rather than words. While perplexity is the randomness or complexity of the word usage, burstiness is the variance of the sentences: their lengths, structures, and tempos. This is a useful point for automatic writing detection - real people tend to write in different sentence formats (through bursts and lulls), often switching structures up. Depending on the amount of interest in the topic, they write long or short sentences often interchangeably, driven by their own verbal momentum (Kuribayashi et al., 2021). As Doyle and Elkan (2009) note, real texts systematically exhibit the phenomenon of burstiness: a word is more likely to occur again in a document if it has already appeared in the current text.

As a result of a statistical analysis of a multi-language corpus of physics-related texts, Constantoudis et al. (2015) demonstrated how the burstiness of long word appearances contributes more to the language-specific aspects of full-word-length correlations. The authors concluded that the correlations between inter-long word distances are less sensitive to language dependencies. A large-scale parallel study of the same content corpora in 10 languages demonstrated that the universal features are linked more to the correlations of the inter-long word distances while the language-specific aspects are related more to their distributions.

Perplexity, on the other hand, measures how well a language model predicts the next word in a sequence or indicates the uncertainty or unpredictability of the text. In the context of AI and human writing, high perplexity means the text is more unpredictable and diverse, while low perplexity indicates a more predictable and repetitive text (Alexander, 2023). Perplexity is a measure used to evaluate the performance of language models, which indicates how well a model can predict the next word in a sequence of words (Jasper Al Whisperer, 2023). It is a statistical estimation that quantifies how "surprised" the model is when it sees new word data

(Techslang, 2022). It corresponds to "the inverse geometric mean of the joint probability of words in a held-out test corpus C" (Miaschi et al., 2020). Since it is the primary measure of the quality of natural language models, it has been used, among others, to distinguish between formal and colloquial tweets (Gonzalez, 2015), detect the boundaries between varieties belonging to the same language family (Gamallo et al., 2017) or identify speech samples produced by subjects with dementia (Cohen and Pakhomov, 2020) or Specific Language Impairment (Gabani et al., 2009). Its advantages are that it is fast to calculate based on the average log-likelihood of the dataset, useful for estimating the model's uncertainty and statistically robust. However, on the downside, it is not accurate for final evaluation, favouring models trained on outdated datasets and not suitable for making comparisons between datasets (Techslang, 2022).

As large language models are trained with prompts to improve their performance, the quality of a prompt affects the extent to which the model is familiar with the language it contains. As Gonen et al. (2022) show, the lower the perplexity of the prompt, the better the prompt can perform the task. For language teachers and students, it might be particularly interesting to compare human-created vs. automatically-created prompts (Table 1) as well as to see the measures of their statistically-performed perplexity (Table 2).

Table 1. Human-created vs. automatically-created prompts for Large Language Models (Gonen et al., 2022)

All manually created prompts	Examples of similar automatically created prompts
What label best describes this news article?	What's the most accurate label for this news article?
What is this piece of news regarding?	What does this piece of news concern?
Which newspaper section would this article likely appear in?	In what section of the newspaper could this article be published?
What topic is this news article about?	What category does this article fall into?

Table 2. Statistically performed perplexity analysis for sentences of similar meaning (Gonen et al., 2022)

Prompt	Ppl
The following two words are antonyms: "good" and "	10.24
The antonym of the word "good" is "	10.32
The word that has the opposite meaning of the word "good" is "	10.43
The word "good" is the antithesis of the word "	10.85
The word "good" is the opposite of the word "	11.15

These examples indicate how perplexity rises with the use of less predictable items ("antithesis") or less specialised terms (opposite instead of antonym in "The word … is the opposite").

Even though studies that investigate the application of perplexity in measuring the similarity of sets of texts are rather scarce, there are some promising findings in the literature. One such study, McFarlane et al. (2009), compared the general news corpus to the obesity news one, with the conclusion that perplexity increased as content became more general relative to obesity news coverage. This indicates that since statistical language model perplexity can measure the similarity of news content to obesity news coverage, it can be used as the basis for an automated health news topic classifier.

Distinguishing human and AI-generated writing

The two concepts of perplexity and burstiness enable a useful distinction, though not necessarily an easy guess, of human vs. AI authorship of texts (Alexander, 2023):

- AI-generated text may have lower burstiness than human writing because AI models are trained on large datasets and tend to use more common words and phrases. Human writers, however, may use rare words and phrases more often due to their creativity and individuality.
- AI-generated text may have lower perplexity than human writing because AI models are optimized to minimize perplexity during training. This means they tend to generate more predictable text. Human writers, with their complex thought processes and personal experiences, can produce more diverse and less predictable text.

Even though perplexity is generally used for distinguishing human vs. AI output, Kuribayashi et al. (2021) show a case against it, proving that the overall "the lower the perplexity, the more human-like the model" generalization does not stand the test in all languages. Since Japanese has typologically different structures from English, the measure of perplexity is not equally effective in both.

Rather than looking out for language calques or stylistic faults, which was the reality of texts generated by early machine translation, teachers need to look at the unpredictability of content and rarity of words as distinguishing factors. This is a crucial change of expectations that needs to be built throughout teacher training.

A proliferation of AI tools and public interest in the potential of Artificial Intelligence for all areas of life, including education, has sparked the emergence of AI detection tools. Some authors (e.g., ChatGPTZero's creator, Edward Tian), claim that AI content detection is an arms race with AI tools development, where detection tools will be only a step behind, responding to human-made exploits, as for AI detection to be effective it needs constant monitoring, iterating and updating by humans (Tian, 2023). The task becomes even more difficult as writers may use AI tools such as Chat-GPT to make their AI-generated content less detectable, by, for instance, asking AI output to be rewritten in a more human-like fashion, more like a 23-year-old or more like a foreign language learner (Baek, 2023). When AI-generated content is combined with human-written text, it becomes even harder, if not impossible, to detect.

The limitations of certain AI models can help in content detection once one becomes aware of them. For instance, as noted by Gillham (2023a), the widely known GPT-4 model is not exempt from inherent faults, which can be possibly used as guidelines for content verification:

- Hallucination: the model tends to create "hallucinated" facts and reasoning errors, similar to its predecessors.
- Limited knowledge: GPT-4's knowledge is restricted to events before September 2021, thus, reasoning errors and the acceptance of incorrect statements as facts can take place.
- Security concerns: since security concerns are inherent within the code, the model may come up with confident predictions that prove false and it cannot validate works for accuracy and correctness.

While several AI detection tools can be applied also in a cross-checking way, as was done by Gillham 2023a, human judgement, instinct, knowledge of typical errors and calques at particular levels, awareness of "sophistication of structure" and "surprise by content" all lead to an inexplicable assumption that only human foreign language teachers will be able to detect AI-generated content (Marr, 2023). By all means, the sample AI detection tools given below should not be treated in a similar way as plagiarism detection instruments, since checking for AI authorship is a much more nuanced activity than simply checking text for reappearance. The results are much more prone to interpretation than unanimous judgement (Gillham, 2023b):

- If an article is 5% Plagiarized that means 5% of the article text is likely copied from another source.
- If an article has an AI score of 5% and a Human score of 95% there is a 95% chance that the article was human-generated (NOT that 5% of the article is AI-generated).

What follows is a list of GPT-4 content detection tools, which, as recommended by Gillham (2023b), should be cross-applied to determine the authorship of texts:

- Originality.ai, https://originality.ai/
- Writer.com, https://writer.com/
- Copyleaks, https://copyleaks.com/
- SEO.ai, https://seo.ai/
- GPTZero, https://gptzero.me/
- Open.ai, https://openai-openai-detector--qz8sj.hf.space/

However, as regards human detection of AI-generated output, a recent study (Alexander, Savvidou and Alexander, 2023) proves that, when untrained in the procedures of AI assistance and when unaware of its possibilities and shortcomings, both teachers and learners are rather weak at recognizing AI-generated texts. In Alexander et al.'s study, the participants tended to exploit a deficit model of assessment that focuses on error as an indicator of learner writing output, with high levels of technical and grammatical accuracy and sophisticated language use as indicators of AI-generated text. This only shows how important it is for advanced students of English to become aware of the concepts of burstiness and perplexity as measures to effectively recognize artificially-generated output, which is going to be explored in the current study.

Method

The aim of the study

The present study was another exploration of the topic of assisting advanced students' writing with Artificial Intelligence-enhanced tools set forward by our earlier research (Krajka, Olszak, in print). The current study focused on linguistic features of human vs. AI-created writing, the opportunities for distinguishing human and AI output, students' awareness of characteristics of their writing and their actual use of opportunities created by contemporary Artificial Intelligence tools. In particular, the current research strived to answer the following questions:

1. How well do advanced students distinguish human-generated from AI-generated essays? What criteria do they think are decisive about Artificial Intelligence authorship?
2. When trained in the use of an AI-assisted word processor, how do advanced students make use of summarising, generating text and writing from prompts?
3. How do students' essays differ linguistically after training in the use of AI tools from those written before treatment? Do they show differences in burstiness and perplexity?

Participants and the teaching context

The study was conducted between March and June 2023 in an undergraduate applied linguistics study programme at a middle-sized public university in Poland. The students took the double-language (English+Portuguese) teacher and translator training programme, in their second year towards a B.A. degree in both languages. Due to the study organisation, it was not possible to randomise group assignment,

hence, an intact group had to serve as the experimental group. Most participants were Polish (12), a few Ukrainian (3), mostly female (9), with 6 males, all aged 21-22.

As was evidenced by the pre-treatment survey, the participants had very little knowledge of and familiarity with Artificial Intelligence and the tools that use AI algorithms to support text processing and production. They generally believe that automatic translation tools are useful for learners as they expose problems of word-for-word translation (47%, 7 out of 15 students) and foreign language students should be taught how to make good use of AI tools in the writing process (60%, 9 out of 15 students). Moreover, a significant part of the participating students (54%, 8 out of 15 students) believe there is a place for Artificial Intelligence tools such as ChatGPT in learning how to write in a foreign language and AI tools are useful in learning paraphrasing for 40% of participants (6 out of 15 students).

Those positive opinions and expectations about Artificial Intelligence tools and Large Language Models are in sharp contrast with previous experiences with such tools. A significant percentage of the survey participants claim that they are unfamiliar with tools like language corpora (e.g., COCA or BNC) – 74% (11 out of 15 students), AI-assisted word processors (e.g., Lex.page) – 87% (13 out of 15 students), AI-assisted summarising tools (e.g., chatDOC) – 87% (13 out of 15 students), text-to-speech synthesisers (e.g, Ivona or Dragon) – 54% (8 out of 15 students) or computer-assisted translation tools (e.g., MemoQ or Trados) – 54% (8 out of 15 students). Few students admit that they only heard of text-processing tools, like language corpora (e.g., COCA or BNC) – 20% (3 out of 15 respondents), chatbots (e.g., ChatGPT) – 20% (3 out of 15 respondents), speech-to-text transcribers (e.g., Google Cloud) – 53% (8 out of 15 students) and computer-assisted translation tools (e.g., MemoQ or Trados) – 20% (3 out of 15 respondents). Even though almost half of the students have tried automatic translation and parallel text retrieval (e.g., with Linguee or Glosbe), asking chatbots (e.g., ChatGPT) – 60% (9 out of 15 students) and synthesising text to speech (e.g., Ivona or Dragon) – 33% (5 out of 15 students), these attempts were deemed largely unsuccessful. It became evident that the use of Google Translate automatic translation tool (81%, 12 out of 15 students), with all its limitations and inaccuracies, shaped the students' attitudes and expectations towards computer-assisted writing.

Design and procedure

The students were approached by one of the researchers (their regular writing instructor) about the possibility of enhancing their academic writing course with Artificial Intelligence tools. The concept of AI-assisted writing was presented and the students were assured of the potential benefits they may gain from the participation in the experimental treatment. Most importantly, it was made clear that the use of AI tools would not lead to deterioration of their final grades, and that the major benefit of the participation, apart from enhanced skills and strategies for assisting writing, would be unlimited access to Lex.page, an online word processor enhanced with Artificial Intelligence algorithms. All the students from the selected group agreed to participate in the study, assured of their right to withdraw and return to regular pen-and-paper writing instruction whenever desired.

The quasi-experimental treatment took place in the face-to-face weekly classes between the end of March and the beginning of June 2023. The learning environment was composed of the face-to-face component, during which tasks were mainly done offline, with the instructor presenting input materials, sample texts or task solutions with the projector, and the computer-based component, during which the participants interacted on their own with different AI-assisted tools as well as with other students in collaborative tasks.

The quasi-experimental treatment was composed of the regular face-to-face component and the individual online work, and was to enhance the previously planned

writing instruction focused on practising the genre of a report in its different forms. In consecutive weeks, students were gradually introduced to the use of AI in writing through the following steps:

1. Week 1 – distinguishing AI-generated from human writing
2. Week 2 - using AI to paraphrase difficult parts of scientific text
3. Week 3 - using AI to generate answers based on different (unknown and unauthorised) sources
4. Week 4 – interacting with ChatDOC
5. Week 5 – composing text with the help of AI using Lex.page
6. Week 6 – collaborating with AI and with other students via Lex.page
7. Week 7 – evaluation and discussion.

The instrumentation for the study was composed of an attitude survey implemented before and after treatment as well as experimental lesson plans. The writing pieces produced by the participants were evaluated according to the criteria adopted at the beginning of the course. The researchers intended to introduce the new writing environment based on Artificial Intelligence in a relatively unobtrusive way, without changing the existing hierarchy of writing objectives.

Artificial Intelligence tools used for the current study

The beginning of the year 2023 has seen great interest in the capabilities of OpenAI's Chat-GPT, with more and more Artificial Intelligence applications emerging in all areas of life. In the current study, it was an important concern to enable all participants equal access to all the tools, hence, the researchers made every effort possible to obtain free-of-charge licences for participants whenever necessary or to use reliable and stable online tools as instruments. The major Artificial Intelligence application used for the study was Lex.page (https://lex.page/), a word processor joining the standard document creation and editing features with functionalities of collaborative writing and Artificial Intelligence operations of continuing writing, generating text from prompts, getting AI feedback on one's writing, and asking AI to insert a random word. Lex.page is generally available as a paid service, however, the researchers managed to obtain free fully-functional licences for the participating students. Moreover, less complex and freely available services as listed below were selected for the introductory classes to give students confidence in the use of AI tools without overwhelming them with new and robust interfaces:

1. Perplexity (https://www.perplexity.ai/) – an online service generating answers based on different (known and unknown) sources.
2. Explainpaper (https://www.explainpaper.com/dashboard) – an online tool allowing uploaded texts to be paraphrased or asked questions about.
3. ChatDOC (https://chatdoc.com/) – an application enabling "chatting with documents", i.e., a file-based reading assistant that can extract, locate and summarise information from documents.

Results and findings

Summary of essay evaluations by participants

The results of the detailed analysis of the four texts subject to the study are as follows:

Table 3. Results of students' essays evaluation

Participant	Text 1. 100% AI-enhanced	Text 2. 100% AI-enhanced based on human generated outline	Text 3. Mix (AI-enhanced & human-generated)	Text.4 100% human-generated
F1	AI	Mix	Mix	AI
F2	AI	Mix	Student	Mix
F3	AI	AI	Student	AI based on human outline
F4	Student	Mix	AI	Mix
F5	AI	Student	Student	AI based on human outline
F6	AI	AI	Student	Student
F7	Student	AI based on human outline	Student	Mix
F8	AI	Student	Student	Student
F9	Mix	Student	Mix	Mix
M1	AI	AI based on human outline	Student	Mix
M2	AI	Student	Student	Mix
M3	AI based on human outline	AI based on human outline	Student	AI based on human outline
M4	AI based on human outline	AI based on human outline	Student	Mix
M5	AI	AI based on human outline	Student	AI based on human outline
M6	AI	AI based on human outline	Student	Mix

Text 1. An essay fully generated by ChatGPT

When assessing the first sample which was fully generated by ChatGPT, 67% (10 out of 15 participants) stated that the text was fully AI-generated, 13% (2 out of 5 participants) assumed it was written by a student, 13% (2 out of 15 participants) believed it was AI-generated based on human outline, and just one participant stated that it was partially generated by the AI and human. The features that made students believe the essay was fully AI-written were the artificiality of the text (*the text does not sound natural*; *the text is written in unnatural language*; *the text gives random information*), inadequate formatting (*the text does not have separate arguments, just an introduction and extended main body and a conclusion*; *the headline is written in big letters*; *the text gives the impression of being written by a student*), repetitions (*a lot of words/phrases are very repetitive*; *paragraphs repeat the same information but in other words*), and exclusive use of online sources (*the text does not contain too many sources*; *the arguments are taken from the source*; *the text seems to be unbiased by giving contrasting arguments*). The results of the study confirm to some extent the findings of Alexander et al. (2023) and Kuribayashi et al. (2021) who stated that AI-powered texts are more predictable (low perplexity) and use common words and phrases (low burstiness). In summary, most of the respondents highlighted that the text lacks verbal complexity (perplexity), which is a typical feature of AI. The overall results indicate that students were quite successful at distinguishing AI and human-generated writing, with quite a high accuracy rate – 67% (10 out of 15 students). However, as indicated by the quotes above, the reasons given by participants were not related to the concepts of burstiness and perplexity, thus, they might have been random guesses.

Text 2. An essay fully generated by ChatGPT based on a human-generated outline

Almost half of the participants classified the test as AI-enhanced based on a human outline, 20% (3 out of 15 students) regarded the text as both AI-enhanced and human-generated, and 27% (4 out of 15 students) believed it was human-generated. The features which classified the text as fully generated by ChatGPT based on human-generated outline were: improper layout of the text (*the text is unwell organized; the text lacks subheadings; the text has limited amount of linking words, the text is not properly divided; the text contains only definitions without author's opinions; the text gives the impression that every next word is generated based on the previous one*), and excessive number of repetitions (*there are a lot of repetitions*). However, the features that made respondents believe that the text was human-written were: the chaotic layout of the text (*the text is chaotic in its structure; the text contains too much statistical data),* poor verbal complexity (*the text is chaotic*) as well as ambiguity of verbal fluency (*the text is written in a mechanical language; the arguments in the text are given in a general manner without specific concentration*). Generally, as the results prove, the text contains several uniform sentences, which indicates low burstiness of the text and is connected to common features of AI-enhanced texts, which corresponds to the study results by Kuribayashi et al. (2021) and Doyle and Elkan (2009).

Text 3. An essay generated partly by a human and partly by ChatGPT

The overwhelming majority of respondents assessed the text as fully written by a human author (80%, 12 out of 15 respondents), only 2 people thought that it was generated partly by a human and partly by ChatGPT and only 1 regarded it as fully AI-generated. The study participants who were more inclined to think that the paper was a fully human work based their decision on the proper layout of the text (*the text is properly organised; the text has naturally long sentences; the text has proper and clear subheadings; the text properly divided into introduction, main body and conclusion; the text is coherent and cohesive*), the naturalness of the text (*the text sounds natural; the text takes a multi-level approach to the problem – topic fully explained*), and high verbal complexity (*the text has a lack of repetitions*; *the text is complex and gives a whole spectrum of information*; *it seems that the author reviewed the sources and analysed the information deeply before writing the text*). The study results confirm the findings by Tian (2023) and Baek (2023) who reported that most of the study respondents regarded sentence variety (burstiness) as connected with human writing, whilst repetitions, lack of verbal complexity (perplexity) appearing in the text were regarded as AI features.

Text 4. A human-generated essay

53% (8 out of 15 participants) classified the text as a mix of AI- and human-generated. Out of the remaining 7 respondents, 4 felt that the essay was AI-enhanced based on a human outline, 2 thought the text was fully written by a human and only 1 person believed it was fully AI-powered. According to study participants, the main features that classified the text as human-written were: the proper layout of the text (*some paragraphs have a natural format*; *the text is properly divided into thematic parts*) as well as cohesion and coherence (*the structure is perfect, all the subheadings are proper and provided; the text has proper structure, normal subheadings*). On the other hand, AI-enhanced characteristics (low burstiness and low perplexity) were repetitions (*the text often contains repetitions or pieces of text that were not paragraphed; the sentences in the text are not complex*) and unnaturalness (*the text reads not natural*; *the conclusion seems to be unnatural*). The overall results indicate that study respondents had problems in identifying levels of burstiness and perplexity

of the given texts, which constitute the main characteristics of the AI- and human-generated texts.

In terms of answering the third research question, that is "How do students' essays differ linguistically after training in the use of AI tools from those written before treatment? Do they show differences in burstiness and perplexity?" the study results can be seen in Table 4 below.

Table 4. Summary of the differences in the students' texts

Category	Linguistic complexity before treatment (human writing)	Linguistic complexity after treatment (AI-generated writing)
	Textual richness	
Burstiness	- long, complex sentences *(e.g. Even though some offers might be dedicated only to the particular group of people, they still attract a certain percentage of those interviewed, which is shown by the fact that over one quarter represents adventure seekers who are willing to go on a camping no matter how conditions of accommodation will look like) – high burstiness.* -a number of rare words or phrases *(providing alimentation; availed) – high burstiness.*	- long, complex sentences *(e.g. By expanding, we can meet the needs of our current and potential customers, while also positioning ourselves as a leader in the market for vegetarian catering services) – high burstiness* - common, repetitive words *(e.g. plans; planned; has, customers)* – low burstiness
Perplexity	- text more unpredictable and diverse (e.g. *As if consumer satisfaction is concerned, a large majority of the people questioned expressed their contentment with the current options…*) – high perplexity	-short, not complex sentences (e.g. *This is both inefficient and costly. To overcome this issue, they have some plans. – low perplexity*

The study results reveal some linguistic differences in students' reports before and after the treatment. However, the marked variations are rather spontaneous and not obvious. In some AI-enhanced reports, there are even some features of human writing, like high burstiness.

As regards specific linguistic differences between students' written assignments before and after the treatment, the four aspects of sentence length, sentence complexity, word frequency and the ratio of unique words inserted in the texts were used as objective measures when comparing students' texts. Generally, when analysing the sentence length of the students' reports after the quasi-experimental treatment, most of the study participants formed slightly longer sentences in the post-texts (67%, 10 out of 15 students), while only some students shortened their sentences (33%, 5 out of 15 students). It can be assumed that the application of the AI tools in the writing practice of the quasi-experimental group of students helped to improve and enrich students' writing skills in general.

Quite interestingly, when comparing the pre- and post-reports with the students' AI-enhanced reports, the majority of the students' AI-assisted texts were slightly shorter, that is up to 4-5 words in one sentence (60%, 9 out of 15 students), whilst some students (40%, 6 out of 15 involved) created a bit longer sentences - up to 5-6 words.

When comparing sentence complexity of the students' pre- and post-reports, slightly more than half of the respondents (53%, 8 out of the involved) created more complex sentences in the post-texts. More specifically, on average 55% of the sentences were compound sentences, 22% constituted complex sentences and 8% of the sentences were compound-complex ones. In contrast to the pre-texts in the post-tests, the compound sentences constituted 53%, the complex sentences 18% and 9% were the compound-complex ones. Quite surprisingly, the analysis of the sentence complexity in the AI-enhanced texts shows that 62% of the sentences were compound, 10% complex, 7% compound-complex and almost 26% were simple sentences comprising one independent clause. Thus, the results prove that AI-generated texts have lower

perplexity than human writing. This means they tend to generate more predictable text. Human writers, with their complex thought processes and personal experiences, can produce more diverse and less predictable texts, as advocated by Jasper Al Whisperer (2023) and Alexander (2023).

In terms of the third and fourth factors, that is word frequency and the ratio of unique words, the results show that on average the pre-texts consisted of more words than the post-texts (about 90 words more), whilst the AI-enhanced ones were significantly shorter in contrast to the pre- and post-texts of the study participants (approximately 800 words). As far as the ratio of the unique words inserted in the written assignments is concerned, there was a slight difference between the pre-texts (55% of the unique words) and the post-texts (53% of the unique words). However, the analysis of the students' texts written with AI-powered tools indicates a moderately lower amount (49%) of the unique words incorporated in the texts. This proves that AI-generated texts seem to have lower burstiness in comparison to human writing.

Conclusion

There is little doubt that the fields of language teaching and learning could be significantly enhanced by the ongoing development of Artificial Intelligence capabilities. According to the study's findings, the students were not very successful at distinguishing between AL and human-generated writing with accuracy rates between 40-80%. They tended to assess human writing based on technical accuracy and error and AI writing based on verbal complexity. Applying AI-powered tools could help students improve their writing competence and bring about changes to the way academic writing is taught as a whole. However, as the study's findings show, they can also hinder the process of improving writing abilities because they lack imagination, use straightforward and sometimes repetitious vocabulary, respond slowly, or make minor grammatical errors.

The research reveals that even highly proficient, almost bilingual, academic students encounter considerable difficulties when attempting to differentiate human writing from AI-assisted written texts. These challenges arise when analyzing formal and informal texts. To validate these findings and potentially discover suggestions and pedagogical implications for foreign language instructors, it is strongly advised to undertake the study on a larger population of bilingual individuals.

However, the study possesses certain limitations. Credible threats to internal validity, namely history, maturation, and testing effects, cannot be dismissed as they present plausible alternative explanations that have impacted students' linguistic development. Furthermore, limitations regarding external validity are evident as the study involved a small sample of students from a singular department within only one university. Consequently, the outcomes lack generalizability not only to a broader population but even to the university population itself due to the aforementioned reason. Despite these constraints, the study does contribute to the existing corpus of research knowledge and offers useful implications for students, educators, and curriculum developers.

Acknowledgement

The authors would like to express their gratitude to the students of applied linguistics of Maria Curie-Sklodowska University, who kindly agreed to participate in the study, as well as Every Writer Collective, which granted study participants free access to Lex.page. The authors are also grateful to anonymous reviewers, whose insightful comments helped give the present paper its current shape.

Bibliographic references

Alexander, C. (2023). *Best Practices for Using ChatGPT at the University of Nicosia*. Training delivered to University of Nicosia faculty, Nicosia, February 2023.

Alexander, K., Savvidou, Ch., Alexander, Ch. (2023). Who Wrote This Essay? Detecting AI-generated Writing in Second Language Education in Higher Education. *Teaching English with Technology, 23*(2), 25-43. https://doi.org/10.56297/BUKA4060/XHLD5365

Baek, D. H. (2023). *ChatGPT Detector – 11 Tools and how to Get around Detection*. Retrieved July 27, 2023, from SEO.ai, https://seo.ai/blog/chatgpt-detector-tools

Baker, T., & Smith, L. (2019). Educ-AI-tion Rebooted? Exploring the Future of Artificial Intelligence in Schools and Colleges. from Nesta Foundation, https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf

Bonner, E., Lege, R., & Frazier, E. (2023). Large Language Model-based Artificial Intelligence in the Language Classroom: Practical Ideas for Teaching. *Teaching English with Technology, 23*(1), 23-41. https://doi.org/10.56297/BKAM1691/WIEO1749

Cohen, T., & Pakhomov, S. (2020). A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1946-1957. from https://aclanthology.org/2020.acl-main.176v2.pdf.

Constantoudis, V., Kalimeri, M., Diakonos, F., Karamanos, K., & Papageorgiou, H. (2015). Long-range Correlations and Burstiness in Written Texts: Universal and Language-specific Aspects. *International Journal of Modern Physics B, 29*, 1541005 (1-3). Retrieved July 27, 2023, from https://doi.org/10.1142/S0217979215410052.

Doyle, G., & Elkan, C. (2009). Accounting for Burstiness in Topic Models. *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 281-288. from https://doi.org/10.1145/1553374.1553410

Frazier, E., Bonner, E., & Lege, R. (2020). Creating Custom AI Applications for Student-oriented Conversations. The FLTMAG. Retrieved July 27, 2023, from https://fltmag.com/creating-custom-ai-applications-for-student-oriented-conversations/.

Gabani, K., Sherman, M., Solorio, T., Liu, Y., Bedore, L., & Pena, E. (2009). A Corpus-Based Approach for the Prediction of Language Impairment in Monolingual English and Spanish-English Bilingual Children. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics. 46-55. Retrieved July 27, 2023, from https://aclanthology.org/N09-1006.pdf.

Gamallo, P., Ramom Pichel, J., & Alegria, I. (2017). A perplexity-based method for similar languages discrimination. VarDial 2017 workshop at EACL2017. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*. Valencia, Spain, 109-114, Retrieved July 27, 2023, from https://aclanthology.org/W17-1213.pdf.

Gillham, J. (2023a) *How to Detect GPT-4 Content*. Originality.ai, Retrieved July 27, 2023, from https://originality.ai/how-to-detect-gpt-4-content/

Gillham, J. (2023b). *AI vs Human Content Detection Score Meaning. Originality*.AI blog, Retrieved July 27, 2023, from https://originality.ai/ai-vs-human-content-detection-score-meaning/

Gonen, H., Iyer, S, Blevins, T., Smith, N. A., & Zettlemoyer, L. (2022). *Demystifying Prompts in Language Models via Perplexity* Estimation. DOI: arXiv:2212.04037v1 [cs.CL].

Gonzalez, M. (2015). An Analysis of Twitter Corpora and the Differences between Formal and Colloquial Tweets. *Proceedings of the Tweet Translation Workshop 2015 co-located with 31st Conference of the Spanish Society for Natural Language*

Processing (SEPLN 2015), Alicante, Spain, Retrieved July 27, 2023, from https://ceur-ws.org/Vol-1445/tweetmt-1-gonzalez.pdf
Jasper Al Whisperer (2023). *The Dummy Guide to 'Perplexity' and 'Burstiness' in AI-generated Content*. Retrieved July 27, 2023, from https://medium.com/@TheJasperWhisperer/the-dummy-guide-to-perplexity-and-burstiness-in-ai-generated-content-1b4cb31e5a81
Kasneci, E. et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences, 103*, 102274. Retrieved July 27, 2023, from https://doi.org/10.1016/j.lindif.2023.102274
Krajka, J., & Olszak, I. (in print). Artificial Intelligence Tools in Academic Writing Instruction – Exploring the Potential of On-Demand Assistance in the Writing Process. Submitted for publication to Roczniki Humanistyczne.
Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower Perplexity is not Always Human-like. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Paper*s), DOI: 10.18653/v1/2021.acl-long
Luckin, R., Holmes, W., Griffiths, M. & Forcier, L. B. (2016). *Intelligence Unleashed. An Argument for AI in Education*. London: Pearson. ISBN 9780992424886
Marr, B. (2023). *How to Detect if Content was Created by ChatGPT and Other Aids*. Forbes.com, Retrieved July 27, 2023, from https://www.forbes.com/sites/bernardmarr/2023/05/25/how-can-you-detect-if-content-was-created-by-chatgpt-and-other-ais/?sh=710eaf47710b
McFarlane, D. J., Elhadad, N., & Kukafka, R. (2009). Perplexity Analysis of Obesity News Coverage. *AMIA 2009 Symposium Proceedings*, 426-430. Retrieved July 27, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815490/pdf/amia-f2009-426.pdf.
Miaschi, A., Alzetta, Ch., Brunato, D., Dell'Orletta, F., & Venturi, G. (2020). Is Neural Language Model Perplexity Related to Readability? *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020*). Retrieved July 27, 2023, from http://ceur-ws.org/Vol-2769/paper_57.pdf
Pokrivcakova, S. (2019). Preparing Teachers for the Application of AI-powered Technologies in Foreign Language Education. *Journal of Language and Cultural Education, 7*(3), 135-153. Retrieved July 27, 2023, from https://doi.org/10.2478/jolace-2019-0025
Sejnowski, T. (2023). Large Language Models and the Reverse Turing Test. *Neural Computation, 35*, 309-342. DOI: 10.48550/arXiv.2207.14382
Techslang (2022). *Perplexity in NLP: Definition, Pros and Cons*. Techslang, September 20, Retrieved July 27, 2023, from https://www.techslang.com/perplexity-in-nlp-definition-pros-and-cons/
Tian, E. (2023). *GPTZero Case Study: Models and Exploits*. Retrieved July 27, 2023, from https://gptzero.me/blogs/gptzero-case-study
Zhao, W. X. et al. (2023*). A Survey of Large Language Models (LLMs*). DOI: arXiv:2303.18223v10

Words: 6850
Characters: 47 006 (26,1 standard pages)

Assist. Prof. Jarosław Krajka, Ph.D., Dr. Litt.
Department of Applied Linguistics
Maria Curie-Sklodowska University, Lublin,
Poland
ORCID: 0000-0002-4172-9960.
jaroslaw.krajka@mail.umcs.pl

Assist. Prof. Izabela Olszak, Ph.D.
Department of Applied Linguistics,
John Paul II Catholic University of Lublin
Poland
ORCID: 0000-0002-8504-7814.
izabela.olszak@kul.pl