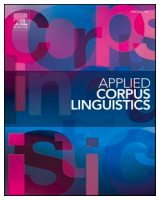
Applied Corpus Linguistics 4 (2024) 100083

Contents lists available at ScienceDirect

Applied Corpus Linguistics

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



Articles

AI-generated vs human-authored texts: A multidimensional comparison

Tony Berber Sardinha *

Graduate Program in Applied Linguistics and Language Studies, Pontifical Catholic University of Sao Paulo, Brazil

ARTICLE INFO

Keywords:
Artificial intelligence
Multidimensional analysis
Register

ABSTRACT

The goal of this study is to assess the degree of resemblance between texts generated by artificial intelligence (GPT) and (written and spoken) texts produced by human individuals in real-world settings. A comparative analysis was conducted along the five main dimensions of variation that Biber (1988) identified. The findings revealed significant disparities between AI-generated and human-authored texts, with the AI-generated texts generally failing to exhibit resemblance to their human counterparts. Furthermore, a linear discriminant analysis, performed to measure the predictive potential of dimension scores for identifying the authorship of texts, demonstrated that AI-generated texts could be identified with relative ease based on their multidimensional profile. Collectively, the results underscore the current limitations of AI text generation in emulating natural human communication. This finding counters popular fears that AI will replace humans in textual communication. Rather, our findings suggest that, at present, AI's ability to capture the intricate patterns of natural language remains limited.

1. Introduction

Since the rise of generative artificial intelligence (AI) chatbots like Chat-GPT, conversations have intensified both in the media and among specialized circles, around the potential replacement of human-generated texts with those generated by AI. This belief is rooted in examples where AI bots have composed specific texts in a particular language that deceive individuals into believing the content to be of human origin. These instances have gained traction, resulting in a backlash against chatbots, with some people even advocating for legislative measures to ban their use (CNBC, 2023).

A current prevailing belief, fueled largely by anecdotal evidence, suggests that AI proficiently substitutes human involvement in language-related tasks. However, an extensive linguistic study that systematically compares AI-generated and human-authored texts across various text varieties is lacking. Such a study would provide insights into the actual proficiency of current AI bots in emulating human language within specific registers.

Given this need, our goal is to assess the human-like quality of texts generated by AI technology from a quantitative multidimensional perspective (Berber Sardinha and Veirano Pinto, 2014; 2019; Biber, 1988, 1995; Friginal and Biber, 2016). We achieve this by comparing texts from distinct registers to texts from corresponding registers authored by humans. Our rationale for employing register distinctions stems from the anticipation that AI-developed chat performance will exhibit variation based on registers. This rationale is consistent with the text-linguistic approach to register variation (Biber, 2019), which posits that language usage is patterned according to register (Biber, 1994, 2019; Biber and Atkinson, 1994).

The current study rests on the assumption that, in order to evaluate the quality of AI-generated texts, one must consider register. Previous corpus-based analysis of human-authored texts has provided ample evidence that register is a key predictor of linguistic variation (Biber, 2012); therefore, studies that seek to determine if human-authored and AI-generated texts vary with respect to each other must take register into consideration for two reasons. First, regular language use is conventionalized and, in real contexts, this corresponds to making language choices that match predefined registers. Thus, the language choices follow certain norms and expectations that are distinct to a particular register. Second, AI users generally ask the AI to produce texts to perform a particular job in a particular context, which requires some form of register awareness. By understanding the norms and expectations of a register, an AI can tailor its language production to better fit the specific context. This adaptation is necessary because the effectiveness of AI-generated texts depends heavily on how well they align with the communicative goals and linguistic conventions of the intended

* Correspondence at: R. Ministro Godoi 969, 4B-02 Perdizes, São Paulo, SP 05015-001, Brazil.
E-mail address: tonycorpuslg@gmail.com.

https://doi.org/10.1016/j.acorp.2023.100083
Received 12 August 2023; Received in revised form 28 November 2023; Accepted 16 December 2023
Available online 20 December 2023

register. If an AI cannot differentiate and adapt to register-specific requirements, the texts it produces will seem out of place or ineffective in achieving their communicative purpose.

The question of whether AI-generated language can accurately reflect register variation remains an underexplored area in literature. Whereas most studies concentrate on the detectability of AI-generated texts compared to human-authored texts, irrespective of register, few investigate its ability to produce responses specific to different registers. For instance, Theocharopoulos et al., (2023) compared human- and AI-generated research article abstracts, employing NLP techniques like word frequency markedness, word embeddings, and named entity recognition, achieving an accuracy of 98.7 % in distinguishing between the two. Some studies have used human judgment rather than automatic measures to detect AI-generated texts, like Ma et al., (2023). These authors, who used ChatGPT to generate research paper abstracts, found that human readers could identify AI-generated texts with high precision (76 %) and recall (75 %) by relying on the fact that the AI struggles to replicate human-created content accurately, particularly when conveying factual information. In contrast, Köbis and Mossink (2021) used ChatGPT to write poetry, which was then compared to work by acclaimed poets. Their study revealed that participants had difficulty reliably detecting AI-generated poetry, a finding that contrasts with their high confidence in identifying artificial poetry and their preference for human-written poetry. Interestingly, these studies suggest that AI may be more adept at emulating human writing in creative registers, such as poetry, than in more formulaic ones like article abstracts, based on human judgment.

For the development of AI, assessing the degree of register awareness in AI models is a key question because it has implications for real-world applications, technological sophistication, and user experience. In practical terms, adaptability to different language registers is essential for the success of AI applications in varied contexts. For example, an AI used in professional environments must be capable of generating a range of professional registers whereas a chatbot designed for social media would need to generate a range of different social media posts. This adaptability is also a measure of model sophistication, indicating not just its ability to generate language, but also its understanding of context, audience, and textual patterns. Moreover, the effectiveness of AI in mimicking human-like register variation significantly impacts user experience. AI that can naturally and appropriately adjust its language to fit a particular register is likely to be more accepted and viewed as proficient.

2. AI register awareness and human-authored text emulation

AI models are often perceived as being able to recognize and generate different registers and styles, as shown by their responses to user requests for texts written in a particular style or resembling a particular text variety. This capability is a key selling point of the models' perceived intelligence. Indeed, various press reports have pointed out that the ability to emulate particular text types makes AI a threat to various professions whose job it is to write specialized texts, such as journalists (Hille, 2023), and even creative artists, such as songwriters (Rascoe and Thompson, 2023).

However, the process by which AI language models learn to differentiate and reproduce various text registers is not clear, mainly because the training methodologies are proprietary and not publicly disclosed.[1] It is believed that each company has developed its own techniques to enable its model to learn register and style distinctions. Nevertheless, it is possible to assume that register awareness comes in part from labeled data during the training phase of the model; in most cases, these are broad labels that reflect the data source. For instance, texts drawn from news sources may be tagged as "news" whereas texts sampled from song lyric websites may be labeled as "songs." On the other hand, for the majority of the training data, which lack explicit labels, the model is designed to deduce the variety from the context, using features like text structure, linguistic elements, and writing style. Either way, some measure of register awareness is sought after and preserved in the model.

The problem of explaining how AI models acquire register awareness is complicated by the fact that, by their very nature, neural networks are considered "black boxes" because their internal decision-making processes are often not transparent or easily interpretable, even to the experts who design them. In general, it is widely accepted that the capacity of neural networks to recognize and duplicate patterns from various registers and styles is primarily a result of the underlying mathematical algorithms and model weights, not of the explicit teaching of the model to mimic these specific forms. This ability emerges organically from the way the model processes and analyzes vast textual corpora rather than from direct instruction in different linguistic styles.

Register awareness is part of the more general goal of AI models to produce human-like text forms, which they learn through a machine learning technique called Generative Adversarial Networks (GANs) (Pan et al., 2019). These consist of two neural networks: the generative network and the discriminative network. The job of the generative network is to create data, while the discriminative network evaluates this data to determine if it resembles artificially generated text or not. The generative network learns to produce data by trying to mimic real, human-generated data. Its goal is to make data indistinguishable from actual human data. The discriminative network acts as a judge. It looks at both real human data and the data created by the generative network, trying to figure out which is which.

As the generative network creates data, the discriminative network evaluates it and gives feedback. If the discriminative network can easily tell that the data are not human-generated, the generative network uses this feedback to improve. Over time, this process leads the generative network to become better at creating data that look like they were made by humans. The ultimate goal of the generative network is to make data that the discriminative network cannot distinguish from real human data. This back-and-forth process makes GANs effective for generating realistic data in AI applications. In addition, GANs are beneficial as they work to reduce the need for extensively annotated training data, streamlining the learning process.

3. Methods

In his 1988 study, Biber detailed how linguistic features in English cluster together to enable users to perform different functions through written and spoken language. He termed these aggregated sets of variables 'dimensions' of variation. Over the years, these dimensions have maintained their relevance and continue to serve as dependable indicators of variation in English (Berber Sardinha and Veirano Pinto, 2014; Friginal, 2013), and as such we rely on them for the purpose of contrasting human-authored and AI-generated texts. Biber (1988) unveiled the following dimensions:

1. Involved versus Informational Production
2. Narrative versus Non-Narrative Concerns
3. Explicit versus Situation-Dependent Reference
4. Overt Expression of Persuasion
5. Abstract versus Non-Abstract Information

These dimensions are based on a corpus comprising 481 texts with approximately 960,000 words (see Table 1), which combines the Brown Corpus (corresponding to the written component, registers 1 to 15), the

[1] According to Kublik and Saboo (2022, pp. 5-6), GPT-3 was trained on five major datasets: Common Crawl, which includes data and metadata from extensive web crawling; Webtext2, based on a large set of web pages; Books1 and Book2, comprising books on a range of subjects; and Wikipedia, incorporating the entirety of the English online encyclopedia.

Table 1
Design of the Biber's (1988) corpus.

	Register	Texts
1	Press reportage	44
2	Editorials	27
3	Press reviews	17
4	Religion	17
5	Skills and hobbies	14
6	Popular lore	14
7	Biographies	14
8	Official documents	14
9	Academic prose	80
10	General fiction	29
11	Mystery fiction	13
12	Science fiction	6
13	Adventure fiction	13
14	Romantic fiction	13
15	Humor	9
16	Personal letters	6
17	Professional letters	10
18	Face-to-face conversations	44
19	Telephone conversations	27
20	Public conversations, debates, and interviews	22
21	Broadcasts	18
22	Spontaneous speeches	16
23	Planned speeches	14

London-Lund Corpus (the spoken component, registers 18–23), and texts collected by the author (registers 16–17), whose composition appears in Table 1.

This corpus formed the foundation for the inaugural Multidimensional Analysis of register variation (Biber, 1988), which aimed to unveil the intricate patterns of variation across texts through the utilization of multiple linguistic features and multivariate statistical techniques. The methodology required tagging each text using the Biber Tagger, which annotated a wide range of linguistic characteristics at both the morphological and syntactic levels. The counts of these linguistic features were subsequently tallied and normed to a rate per 1000 words, facilitating the comparison of texts of differing lengths.

These counts were entered into a factorial analysis, which revealed the groups of correlated linguistic features in the data. The factors were interpreted by means of a careful qualitative analysis to determine the underlying communicative functions performed by the linguistic features loading on each factor and were labeled using a descriptor that reflected its major communicative function in the texts, thereby giving rise to the dimensions of register variation. The texts were scored on each dimension by adding up the (standardized) counts of the features loading on (each pole of) each factor, and mean factor scores were computed for each register, which enabled comparing the registers along each dimension.

Since Biber's 1988 publication, these dimensions have been used in many different studies as points of comparison for understanding register variation in English (e.g.Berber Sardinha and Veirano Pinto, 2019; Conrad and Biber, 2001). Such comparison is possible because researchers calculate factor scores for their respective corpora based on the dimensions of variation from a prior study, a method that has become known as 'additive analysis' (Berber Sardinha et al., 2019). In so doing, the researcher can 'add' their corpus to a previous set of dimensions and describe these registers using those dimensions as a reference. The 1988 dimensions have been widely used as a reference due to their reputation as dependable parameters of variation, thereby serving as 'yardsticks' for measuring registers that were not included in the 1988 study, like digital registers (Berber Sardinha, 2014), or AI-generated texts (as in this study), which had yet to be developed.

Broadly speaking, detecting AI-generated texts from human-authored texts can be framed as detecting non-naturalness or artificiality in language texts—an issue that has motivated previous research in corpus linguistics from a multi-dimensional perspective. The issue of naturalness versus artificiality in communication has been examined in different settings, including television, film, and educational materials. More specifically, the goal has been to determine the extent to which particular registers like television programs, movies, and textbook conversations can successfully emulate naturally occurring conversation. Unlike real conversations, dialog found on television programs, cinematic productions, and educational materials is planned, scripted, rehearsed, and tailored to fit a plot or educational objective, rather than to serve the communicative goals of real speakers engaged in face-to-face interactions. However, despite the different production circumstances, experienced scriptwriters and material designers can compose dialog that passes as authentic conversations. The question then arises as to whether the resulting talk is linguistic similar to actual conversations, despite the sharp contextual differences.

As a result, researchers have resorted to Biber's (1988) dimensions of register variation to assess the naturalness of the conversation found in film (Forchini, 2012; Veirano Pinto, 2013; Veirano Pinto and Forchini, in press), television (Al-Surmi, 2012, 2022; Berber Sardinha and Shimazumi, 2021; Berber Sardinha and Veirano Pinto, 2017; Quaglio, 2009), and textbooks (Le Foll, 2021). Overall, previous studies have suggested that the distribution of linguistic features in television and

Table 2
Linguistic features loading on each dimension (Biber, 1988).

Dimension	Linguistic characteristics
1 Involved vs Informational Production	Involved Production: private verbs, *that* deletion, contractions, present-tense verbs, second person pronouns, *do* as pro-verb, analytic negation, demonstrative pronouns, general emphatics, first person pronouns, pronoun *it*, *be* as main verb, causative subordination, discourse particles, indefinite pronouns, general hedges, amplifiers, sentence relatives, *wh* questions, possibility modals, nonphrasal coordination, *wh* clauses, final prepositions, adverbs Informational production: nouns, word length, prepositions, type/token ratio, attributive adjectives, place adverbials
2 Narrative vs non-narrative concerns	Narrative concerns: past-tense verbs, third person pronouns, perfect-aspect verbs, public verbs, synthetic negation, present-participial clauses Non-narrative concerns: present-tense verbs, attributive adjectives
3 Explicit versus situation-dependent reference	Explicit reference: *wh* relative clauses on object position, pied-piping relative clauses, *wh* relative clauses on subject position, phrasal coordination, nominalizations Situation-dependent Reference: time adverbials, place adverbials, adverbs
4 Overt expression of persuasion	infinitives, prediction modals, suasive verbs, conditional subordination, necessity modals, split auxiliaries, possibility modals
5 Abstract versus non-abstract style	Abstract style: conjuncts, agentless passives, past-participial clauses, *by* passives, past-participial WHIZ deletion, other adverbial subordinators Non-abstract style: No features

film dialog resembles those found in face-to-face conversation. In contrast, research shows that textbook conversations differ considerably from authentic conversations (Le Foll, 2021). Although some textbook conversations can be linguistically close to actual conversations, many include a disproportionate number of linguistic features associated with the literate end of Biber's dimension 1 (such as nouns, prepositions, and attributive adjectives), rendering them artificial in comparison to face-to-face conversations.

Based on the previous literature that relied on the dimensions of register variation from Biber (1988) to measure the degree of naturalness versus artificiality of simulated texts, these dimensions were applied in the current study to gauge the naturalness of AI-generated texts.

The linguistic characteristics comprising each dimension appears in Table 2.

In this study, we conducted an additive analysis using a corpus compiled specifically for this research. We employed the 1988 dimensions of variation as a reference point to explore the human-like attributes of AI-generated texts. Specifically, our aim is to compare AI-generated texts from specific registers against their human-generated counterparts within the same registers. The design of the corpus compiled for this study appears in Table 3.

The AI component of the corpus was generated using ChatGPT, via the API using model 3.5 in July 2023. The API was accessed using a *curl* script that delivered a prompt to the chat and returned its output. These prompts were tailored for each register:

- Conversation: Write a conversation between people of about 1000 words in length.
- Academic: Write an introduction of an article in the field of [Chemistry/Applied Linguistics] totaling about 1000 words in length.
- L2 Essay: Write an argumentative essay of about 400 words in length for a student assignment for an English as a foreign language class, using no headings, in a single paragraph.
- News: Write a news article of about 500 words in length.

The output were JSON files that were processed using a script developed for this study, resulting in plain text files.

The human-authored component of the corpus was drawn from existing corpora, as follows:

- Conversation: The texts were selected in a random manner from the 2014 British National Corpus (BNC; Love et al., 2017). Due to the constraints on ChatGPT output length, using entire conversations was unfeasible as they often extended beyond the chatbot's output capacity. Consequently, an approximate 1000-word excerpt was sampled from each BNC file. A script was developed for the purpose of converting XML files into plain text format and for extracting the subsequent complete turn beyond the 1000-word limit, thereby ensuring the preservation of entire turns.
- Academic: Drawing from prior corpus-based investigations that revealed considerable lexicogrammatical variation across academic disciplines (Gray, 2015; Hardy, 2015) and article sections (Prina Dutra and Berber Sardinha, 2021), this component encompassed two distinct fields: applied linguistics and chemistry. For each article, a single section—the introduction—was included. The data were drawn randomly from the CorAAL and CorAChem corpora compiled by Dutra et al., (2020), which comprise research articles published in major journals in several fields.
- L2 Essay: A random sample of 100 texts was drawn from the International Corpus of Learner English (ICLE; Granger et al., 2020).
- News: A random sample was drawn from the January 2019 holdings of the News on the Web (NOW) corpus, developed and distributed by Davies (2023). This time period was selected because ChatGPT was first released to the public in November 2019 (Vincent, 2023), thereby ensuring that these texts were not aided or crafted by AI.

The corpora used in this study were tagged by the Biber Tagger, whose tagset includes 217 linguistic feature tags. After tagging, the tagged texts were post-processed with the Biber Tag Count, a separate software program that tabulates the tags and reports normed counts per one thousand words for a total of 154 features, including the scores for each individual text on each of the first five dimensions from Biber (1988).

A dimension score can be described as a numerical value that indicates the relative position of a text on a dimension. These scores are used to estimate the degree to which a text possesses or is influenced by the underlying dimension that the factor represents. As the dimensions are not directly observed but are inferred from the patterns of correlations among the observed variables (the linguistic features), the scores allow for the comparison of texts in terms of these underlying dimensions.

For those dimensions comprising two poles (such as dimensions 1, 2, and 3), the dimension scores were computed by adding up the standardized counts (z-scores) of the features loading on the positive pole and then subtracting the sum for the features loading on the negative pole. For the remaining dimensions, the scores were calculated by summing up the standardized counts of the features loading on the positive pole.

The dimension scores were used for two main goals. The first goal was to determine whether it was possible to distinguish between AI-generated and human-authored texts by comparing the mean scores of each authorship condition through analysis of variance (ANOVA) and the coefficient of determination (R^2). ANOVAs allowed for testing the statistical significance of differences between the mean scores of human-authored and AI-generated texts, indicating if these differences were due to chance or if they were statistically significant. Meanwhile, the coefficient of determination provided a measure of how well the variability

Table 3
Design of the corpus used in the study.

Source	Register	Texts	Word Count	Sub-register		Word Count
AI	Academic	100	63,981	Applied Linguistics	50	32,226
				Chemistry	50	31,775
	Conversation	100	82,758			
	L2 Essay	100	49,935			
	News	100	59,945			
		400	256,619			
Human	Academic	100	89,095	Applied Linguistics	50	52,311
				Chemistry	50	36,784
	Conversation	100	90,086			
	L2 Essay	100	60,240			
	News	100	50,528			
		400	289,949			
Total	8	800	546,568			

in the scores could be explained by the fact that the text was human authored or AI generated.

The comparison was conducted using SAS OnDemand, employing the GLM procedure. The analyses were carried out under two distinct conditions: one where register was included as a main effect and another where it was not. In the scenarios where register was not considered a main effect, the ANOVAs and R^2 values primarily focused on determining the degree to which human-authored and AI-generated texts could be differentiated, without taking into account register distinctions. On the other hand, when register was factored in as a main effect, the analyses were geared toward assessing the influence of register in predicting whether a text was authored by a human or an AI. This measured the impact of register on the distinguishability of human-authored and AI-generated texts.

The second goal was to predict authorship—more specifically, to determine the degree to which it was possible to correctly classify the human-authored and AI-generated texts for each register. This was carried out using discriminant function analysis (DFA; see Cantos Gómez, 2013; Crossley et al., 2014; Veirano Pinto, 2019) through PROC DISCRIM in SAS OnDemand. In this DFA, the primary goal was to use a set of predictive variables (the dimension scores) to differentiate between the groups in the data—in this case, human-authored and AI-generated texts, further broken down by register. This approach took into account not only the overarching categories of human versus AI authorship but also how these categories interact with and are influenced by different registers. The procedure involved creating a discriminant function, essentially a linear combination of the dimension scores, which maximized the differences between groups. By applying this function to the data, it categorized each text into the group (a combination of register and authorship) to which it most likely belonged, based on its dimensional characteristics. The efficacy of this classification was then evaluated by considering the percentage of correctly classified cases.

The DFA employed cross-validation, whereby as many predictive models were produced as there were texts in the corpora; for each text whose classification was predicted, the text was removed from the dataset, and the DFA model was trained on the remaining texts. This process was repeated for each text in the corpora, ensuring that each text was used exactly once as a test case. By systematically omitting different texts and testing the predictive accuracy of the model, it was possible to guard against overfitting (where a model is too closely tailored to the specifics of the training data and performs poorly on new data), thereby ensuring that the predictions were not biased by the characteristics of each text being predicted.

In the DFA conducted in our study, misclassifications (i.e., texts not assigned to their expected categories) were relevant because they showed the extent to which AI generation is not register sensitive—that is, the extent to which a text considered by the AI to be from a particular register is linguistically akin to a different register. Register-internal variation is a natural phenomenon because registers are culturally recognized rather than scientifically defined categories (Biber and Egbert, 2023). Given that register boundaries are not defined by linguistic characteristics, a certain degree of misclassification is anticipated in human-generated texts because texts from different registers may sometimes show more linguistic similarities with each other than with texts from their own register; consequently, they can be mistakenly identified as belonging to a different register.

Register-internal variation is a well-recognized phenomenon in human-generated texts, reflecting their origin in human culture. However, such variation within AI-generated texts has not yet been documented. Consequently, it is unclear to what extent AI-generated texts exhibit variability within the register categories they are prompted to emulate. Furthermore, the degree to which this variation in AI texts occurs compared to the levels of register-internal variation observed in human-generated texts is also unknown. This gap in understanding highlights a crucial area of inquiry in evaluating the sophistication and adaptability of AI in mimicking human-like textual variation within specific registers.

Table 4
Mean dimension scores for source.

Dimension		AI	Human	F	R^2
1	Mean	-12.1	1.7	75.51	0.09
	SD	19.7	24.9	$p<0.0001$	
2	Mean	-2.7	1.4	65.56	0.07
	SD	-1.6	2.3	$p<0.0001$	
3	Mean	-0.3	0.1	1.9	2.00
	SD	3.0	5.2	NS	
4	Mean	-2.5	-1.0	47.72	0.06
	SD	2.6	3.8	$p<0.0001$	
5	Mean	2.3	2.0	2.02	2.00
	SD	3.1	4.1	NS	

4. Results

Considering our view on register as a key driver of variation in language use, we did not anticipate that AI-generated and human-authored texts could be differentiated from one another without accounting for register distinctions. To test this assumption, we first compared the AI- and human-generated texts as a whole, without dividing each subcorpus into registers. The results appear in Table 4, which presents the mean scores for human-authored and AI-generated texts as a whole. Without accounting for register distinctions, the average scores between the AI and human production conditions exhibited statistically significant differences for only three out of the five dimensions, and the effect size of these differences (R^2) was quite minimal. The non-significant F-tests and low R^2 values reflect the high standard deviations in both the AI and human data. The wide variability can be attributed to the variation in dimension scores among the registers. This result highlights the critical role of considering registers; without such consideration, one might erroneously conclude that AI and human sources are indistinguishable, which is false, as we will demonstrate by comparing the mean scores for each register.

In contrast to the results observed when register was not treated as an independent variable, the introduction of register reveals a distinct picture, as shown in Table 5, which presents the results of the F-tests and R^2 for each dimension distinguished by register. As Table 5 shows, with the incorporation of register distinctions, all F-tests exhibit high significance, accompanied by notable effect sizes (cf. Goulart and Wood, 2019) across all dimensions.

Register emerges as a powerful predictor of AI performance in emulating human language, with AI introducing major changes in the dimensional profile of the registers compared to human-authored texts. Three dimensions show greater disparity (>50 %) between AI and humans: involved versus informational production (Dim. 1), overt expression of persuasion (Dim. 3), and abstract versus non-abstract information (Dim. 5). We can conclude that the AI models were not properly trained to exhibit these communicative functions consistently in different English registers. The means for each register along each dimension are shown in Fig. 1 through 5.

Regarding Dimension 1 (Fig. 1), the mean scores show sharp differences across conversations, essays, and news. Notably, AI-generated

Table 5
ANOVA and coefficient of determination for source (AI vs. human) and register.

Dim.	F	p	R^2
1	950.4	<0.0001	0.92
2	57.3	<0.0001	0.40
3	139.8	<0.0001	0.61
4	73.6	<0.0001	0.46
5	108.5	<0.0001	0.55

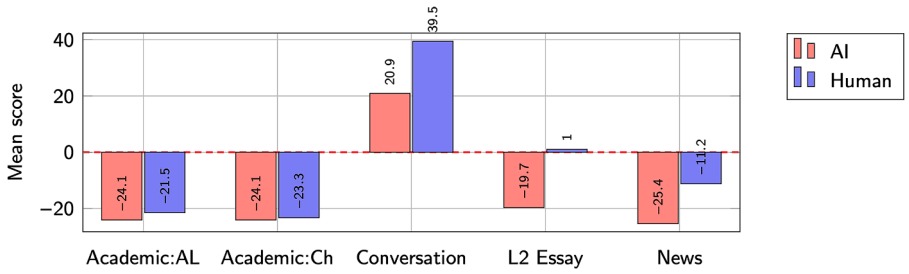


Fig. 1. Mean dimension scores for Dimension 1: Involved versus Informational Production.

conversations tend to exhibit approximately half the degree of involvement features compared to their human-generated counterparts. When compared with Biber's (1988) findings, the mean scores position AI conversations in proximity to personal letters, spontaneous speeches, and interviews but away from face-to-face conversations, whose mean score is 35. Conversely, human conversations from the BNC exhibit a congruence with conversations from the London–Lund Corpus employed by Biber. This discrepancy underscores that AI has not been adequately designed to capture the complexity of spoken language grammar. Instead, it relies on written language patterns to generate spoken dialog.

Examples 1 and 2 contrast typical AI-generated and spontaneous human-generated conversations. Whereas the AI dialog is characterized by a noticeable absence of engagement, the human conversation shows the speakers negotiating meaning through the constant use of linguistic features that permit the expression of personal involvement and interaction. In the AI-generated dialog, these elements are sparse and infrequent, and the dialog fails to encompass the entire range of features typically encountered in a natural face-to-face conversation. The AI-generated text exhibits an artificial quality, resembling more of a question-and-answer or interview session than a conversational exchange; it sounds scripted and lacks the hesitancy, uncertainty, and vagueness commonly found in spontaneous human conversation.

Example 1. Sample from an AI-generated conversation

Speaker A: "Have **you been** on any other international trips in the past?"
Speaker B: "Yes, **I've** had the opportunity to travel to several countries over the years. Some of my favorite destinations **include** France, Italy, and Thailand. Each trip **was** unique **and** offered a chance to explore different cultures, try new foods, **and** visit iconic landmarks. **I believe** traveling **is** a great way to broaden **our** perspectives **and truly appreciate** the world's diversity."
Speaker A: "Do **you** have any other countries on **your** travel bucket list?"
Speaker B: "Absolutely! Japan **is just** the beginning for **me**. Some other destinations on **my** bucket list **include** Australia, Brazil, Egypt, and Iceland. **I want** to witness the Great Barrier Reef, experience Rio de Janeiro's vibrant Carnival, explore the ancient Egyptian pyramids, **and** marvel at Iceland's stunning landscapes. There**'s so** much to see **and** do in **this** world!"

Example 2. Sample from a naturally occurring human conversation (BNC 14 file SPYD.xml)

Speaker A: "**I guess** at the moment **like I'd be so** scared if **I was** the prime minister **right now like**"
Speaker B: "Mm"
Speaker A: "**I** wouldn't **know what** to do"
Speaker B: "Yeah yeah the only positive thing **that's** coming **out** of **this is** the fact that ⟨name⟩ **is now** being tested **which I like a lot**"
Speaker A: "**Like what can you** do in **this** situation? "
Speaker B: "Yeah **right**"
Speaker A: "But **what can you do** in **this** situation? **like** wha **how do you** control"
Speaker B: "**I've** no idea"
Speaker A: "Something like **this**?"
Speaker B: "I **bet** he **doesn't even** do half of **it**"
Speaker A: "Cos **do you think** he **was** expecting **all this** to happen? "

The differences observed in essays and news are even more pronounced. AI essays lean significantly toward information density, whereas human essays are not marked. Furthermore, AI-generated news stories score three times as high on the informational end of the dimension as their human-authored counterparts. In contrast, AI-generated academic texts closely mirror their human-authored counterparts, as evidenced by the scores, which indicate a remarkable degree of similarity. This result highlights the excessive dependence of AI models on academic language, which is edited, formal, and information-packed—characteristics that have been overused in news and essays.

Turning to Dimension 2, the differences between AI- and human-generated texts are less prominent, as indicated by the R^2, which captures the least variation at 40 % (a large amount, nonetheless). As shown in Fig. 2, across all registers, the human-authored texts exhibit a higher degree of narrativity compared to AI-generated texts. Particularly noteworthy are the differences observed in three registers: applied linguistics introductions, conversations, and news. In these registers, the non-narrative character of the texts is notably accentuated in the AI-generated samples.

Dimension 3 accounts for the second-highest level of captured variation, at 61 %. As Fig. 3 indicates, in contrast to the preceding dimensions, the most significant relative divergences emerge within the academic texts. In chemistry, AI texts are approximately 6.5 times less likely to incorporate explicit reference, whereas in applied linguistics, AI

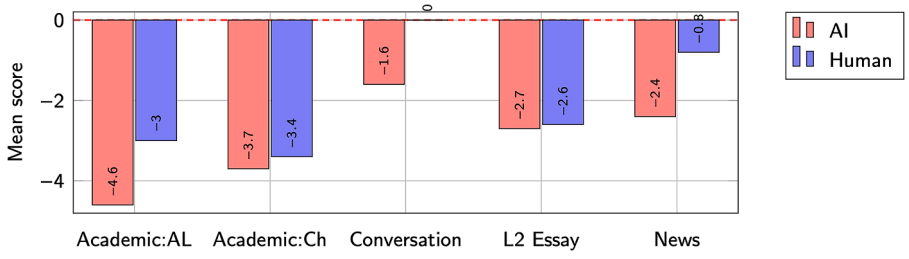


Fig. 2. Mean dimension scores for Dimension 2: Narrative vs Non-narrative Concerns.

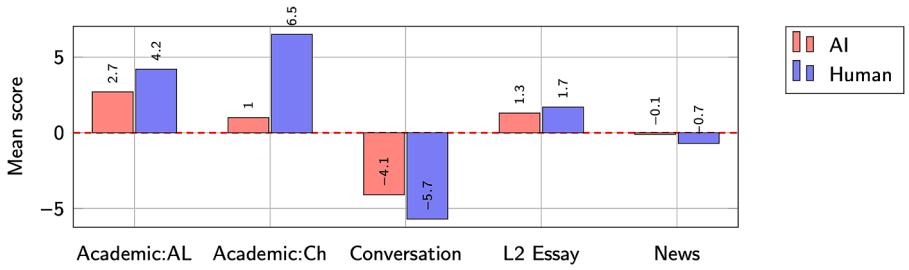


Fig. 3. Mean dimension scores for Dimension 3: Explicit vs Situation-Dependent Reference.

texts are about 1.6 times less likely to use explicit reference compared to human-authored texts. Furthermore, in conversation, AI-generated texts are 1.4 times less likely to employ context-dependent reference. These findings suggest that the AI models do not grasp the intricacies of reference utilization in English and, thus, lack an understanding of the expected incidence of explicit reference and context-dependent reference within specific registers.

Examples 3 and 4 highlight the problems encountered by AI to generate adequate amounts of explicit referencing in texts. The AI sample has an evident deficiency as it includes only two instances of dimensional characteristics (a single instance each of pied-piping construction and nominalization). On the other hand, the human-authored sample incorporates multiple nominalizations followed by *wh-* relative clauses, thereby making the references more explicit.

Example 3. Sample from an AI-generated Chemistry article introduction

Whether studying the vastness of the cosmos or scrutinizing the minute details of a single cell, chemistry is present, interweaving itself in the tapestry of scientific **exploration**. One of the pillars **upon which** modern chemistry stands is the concept of atoms. These minuscule particles, undetectable to the naked eye, are the fundamental units of matter. Overwhelmingly abundant in nature, they combine to form various elements, each possessing unique properties and characteristics.

Example 4. Sample from a Human-authored Chemistry article introduction

A first strategy to reduce CO2 **emissions, which** has been deeply investigated in the last years and **which** has been recently applied for the first time to a large-scale power station in Canada, is Carbon **Capture and Storage** (CCS) [4], **which** consists in the permanent CO2 **storage** deep underground in very specific geological sites. A very attractive alternative to this technology is represented by Carbon **Capture and Utilization** (CCU) processes, **which** consist in the chemical **conversion** of CO2 to added-value carbon-containing products.

On Dimension 4, the disparities between AI-generated and human-authored texts are less pronounced, albeit still substantial, as evidenced by the high captured variation of 46 %. Fig. 4 shows that AI academic texts exhibit a higher degree of non-persuasiveness compared to their human-authored counterparts. Intriguingly, a shift in direction is observed in essays for the first time: AI texts manifest non-persuasiveness whereas human texts adopt a persuasive tone. This dichotomy underscores the AI models' tendency to produce texts that contradict the expected norms in real-world scenarios.

Moving to Dimension 5, we encounter the third most substantial variation between AI-generated and human-authored texts, amounting to 55 %. As depicted in Fig. 5, distinct disparities are evident between AI and human texts, with three registers displaying opposing directions for AI and human output. In the academic context, AI-generated content demonstrates a deficiency in abstract features; meanwhile, human texts normally exhibit significant amounts of abstraction. Contrarily, with respect to conversation, AI dialogs trend toward abstraction in contrast to real-life expectations, where conversations tend to be notably non-abstract. Conversely, essays tend to be generally less abstract compared to the AI-generated texts.

5. Predicting authorship

In this section, we employ the dimension scores as input for a linear discriminant function analysis (Cantos Gómez, 2013; Veirano Pinto, 2019) in an attempt to differentiate and classify texts as either AI-generated or human-authored, based solely on these scores. Initially, a stepwise discriminant analysis was conducted to identify the dimensions most influential in distinguishing between the two authorship conditions. The results indicated that all five dimensions served as strong predictors, prompting us to employ the five dimension scores of each text as variables in the linear discriminant function analysis.

The findings as displayed in Table 6 reveal a 79 % accurate classification rate for AI texts compared to a 69 % accuracy in classification rate for human-authored texts. This result underscores the relative ease with which AI texts can be identified using the dimensions of variation. Intriguingly, the fact that human-generated texts occasionally receive classification as AI-generated texts implies that certain AI-generated texts closely replicate human writing patterns. Consequently, human-generated texts resembling these successfully mimicked AI texts tend to be classified as AI-generated texts.

Table 7 presents a breakdown of the classification results across different registers, offering a view of the extensive variation in

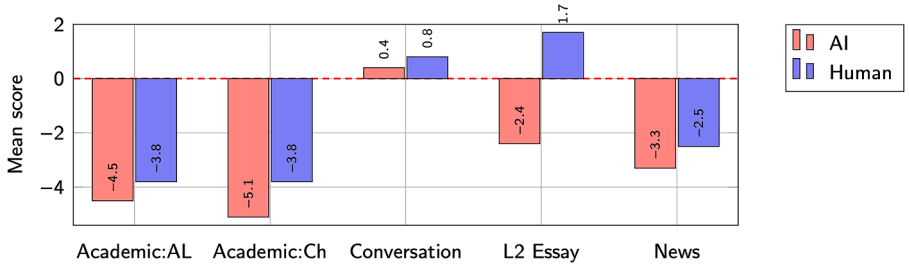


Fig. 4. Mean dimension scores for Dimension 4: Overt expression of persuasion.

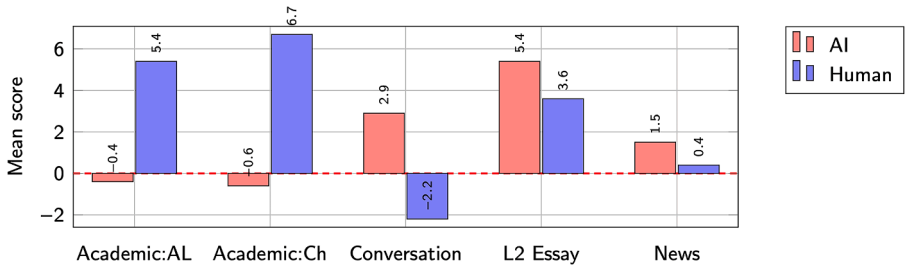


Fig. 5. Mean dimension scores for Dimension 5: Abstract versus non-abstract information.

Table 6
Discriminant function analysis: AI versus human.

	Classified as		
Originally	AI	Human	
AI	316	84	400
	79.00 %	21.00 %	
Human	123	277	400
	30.75 %	69.25 %	
	439	361	
	54.88 %	45.13 %	

classification results. Notably, conversation emerges as the most accurately classified register, achieving an impressive 95 % accuracy for AI and 94 % for human classifications. This result reaffirms the point underscored during the presentation of the individual dimensions: AI struggles to authentically replicate the spontaneous nuances inherent in human conversation. Given that chatbots are designed primarily to engage in human-like conversations, this limitation indicates a significant hurdle, exposing the artificiality of AI dialogs due to their limited grasp of authentic human conversational dynamics.

Among the top four best-classified registers, three are AI-generated texts: conversation, applied linguistics article introductions, and chemistry article introductions. The failure of AI to emulate academic writing is surprising, considering that AI models often draw extensively from academic sources. In contrast, real face-to-face conversation is generally underrepresented in AI training data. This discrepancy in AI performance can likely be attributed to the lack of conversational data in its training set.

In contrast, the four least accurately classified texts consist of three human-authored registers: chemistry introductions, news, and applied linguistics introductions. The misclassification of the two different types of human-sourced article introductions relates to their similar scores across all dimensions. However, unlike articles, human news texts are often misclassified by AI because its training data includes a large amount of news content from websites. This leads AI to produce outputs in various registers that resemble news, resulting in human-generated news being misclassified as any AI-generated register.

The misclassification of human-authored introductions in applied linguistics as AI-generated essays can be attributed to the fact that they share similar dimensional profiles in three out of the five dimensions, specifically Dimensions 2, 3, and 5. However, this partial similarity in profiles is probably due to the AI lacking training on actual student essays, leading it to infer an essay model that coincidentally reflects the dimensionality of human-authored article introductions in Applied Linguistics.

6. Summary and conclusion

The research findings reveal hitherto unknown patterns across multiple dimensions of register variation between AI-generated and human-authored texts. On Dimension 1, AI conversations exhibit

Table 7
Discriminant function analysis: Register breakdown.

			Classified as AI					Classified as human					
	From		AL	Ch	Cnv	Esy	Nws	AL	Ch	Cnv	Esy	Nws	Total
AI	AL	N	36	12	0	0	1	1	0	0	0	0	50
		%	**72**	24	0	0	2	2	0	0	0	0	100
	Ch	N	9	35	0	0	4	2	0	0	0	0	50
		%	18	**70**	0	0	8	4	0	0	0	0	100
	Cnv	N	0	0	95	0	0	0	0	2	3	0	100
		%	0	0	**95**	0	0	0	0	2	3	0	100
	Esy	N	3	3	0	60	12	13	5	0	0	4	100
		%	3	3	0	**60**	12	13	5	0	0	4	100
	Nws	N	1	20	0	9	65	4	0	0	0	1	100
		%	1	20	0	9	**65**	4	0	0	0	1	100
H	AL	N	2	5	0	12	3	7	19	0	1	1	50
		%	4	10	0	24	6	**14**	38	0	2	2	100
	Ch	N	3	0	0	2	0	14	31	0	0	0	50
		%	6	0	0	4	0	28	**62**	0	0	0	100
	Cnv	N	0	0	6	0	0	0	0	94	0	0	100
		%	0	0	6	0	0	0	0	**94**	0	0	100
	Esy	N	0	0	10	8	1	3	4	0	70	4	100
		%	0	0	10	8	1	3	4	0	**70**	4	100
	Nws	N	4	8	4	4	8	1	0	0	11	60	100
		%	4	8	4	4	8	1	0	0	11	**60**	100
	Total	N	58	83	115	95	94	45	59	96	85	70	800
		%	7.25	10.38	14.38	11.88	11.75	5.63	7.38	12	10.63	8.75	100
	Priors	%	10	10	10	10	10	10	10	10	10	10	

reduced involvement and integration compared to their human counterparts, while essays and news stories also demonstrate substantial disparities, with AI essays being notably laden with information-carrying features. Dimension 2 highlights AI's issues with effectively incorporating narrativity, as particularly evident in certain registers like applied linguistics introductions, conversations, and news, where AI-generated texts struggle to integrate narrative elements found in human texts. Dimension 3 reveals deficiencies in how AI generates referential language. In academic texts, AI-generated content lags significantly in incorporating explicit and context-dependent references, a pattern observed in both chemistry and applied linguistics introductions. Dimension 4 also reveals an issue with AI-generated texts. While academic writing and news typically lack persuasive features, AI tends to overdo this aspect, producing texts almost entirely devoid of natural persuasive language. Conversely, in real-life contexts, student essays and conversations often rely on persuasive elements, yet AI-generated texts for these registers continue to lack these features. Finally, with respect to abstraction (Dimension 5), AI academic texts exhibit a shortage of the abstraction indexing features found in human texts. Strikingly, in conversations, AI is on the opposite end of the dimension from human texts, with AI dialog leaning toward greater abstraction, whereas human texts display a tendency toward non-abstraction.

In contrast to the results observed when register was not treated as an independent variable, the introduction of register reveals a distinct picture, as shown in Table 5, which presents the results of the F-tests for each dimension distinguished by register. As Table 5 shows, with the incorporation of register distinctions, all F-tests exhibit high significance, accompanied by notable effect sizes (cf. Goulart and Wood, 2019) across all dimensions.

These findings shed light on the current capabilities of AI models in generating human-like language. Although the AI-generated output might appear convincingly human, subjecting the texts to a rigorous linguistic analysis employing robust measures such as the 1988 dimensions of register variation (Biber, 1988) reveals significant disparities. Profound differences emerge at the lexicogrammatical level of AI texts, underscoring their considerable distance from human-generated texts.

CRediT authorship contribution statement

Tony Berber Sardinha: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author wishes to thank the following organizations for funding the research reported here: Grant #2022/05848–7, São Paulo Research Foundation (FAPESP); Grant #310140/2021–8, CNPq (Brasília, Brazil); Grants #29441, #26726, PIPeq PUCSP.

References

Al-Surmi, M., 2012. Authenticity and TV shows: a multidimensional analysis perspective. TESOL Q. 46 (4), 671–694.
Al-Surmi, M., 2022. TV shows, authenticity, and language learning: a corpus-based case study. Register Stud. 4 (1), 30–54. https://doi.org/10.1075/rs.19016.als.
Berber Sardinha, T., Shimazumi, M., 2021. What's on the telly? A multi-dimensional analysis of register variation in British television. Paper presented online at the 11th International Corpus Linguistics Conference, University of Limerick.
Berber Sardinha, T., Veirano Pinto, M. (Eds.), 2014. Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber. John Benjamins.
Berber Sardinha, T., Veirano Pinto, M., 2017. American television and off-screen registers: a corpus-based comparison. Corpora 12 (1), 85–114.
Berber Sardinha, T., Veirano Pinto, M. (Eds.), 2019. Multi-Dimensional Analysis: Research Methods and Current Issues. Bloomsbury Academic, London.
Berber Sardinha, T., Veirano Pinto, M., Mayer, C., Zuppardi, M.C., Kauffmann, C., 2019. Adding registers to a previous multi-dimensional analysis. Eds. In: Berber Sardinha, T., Veirano Pinto, M. (Eds.), Multi-Dimensional Analysis: Research Methods and Current Issues. Bloomsbury Academic, London, pp. 165–188.
Berber Sardinha, T., 2014. 25 years later: comparing Internet and pre-internet registers. Eds. In: Berber Sardinha, T., Veirano Pinto, M. (Eds.), Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber. John Benjamins, Amsterdam, pp. 81–105.
Biber, D., Atkinson, D., 1994. Register: a review of empirical research. Eds. In: Biber, D., Finegan, E. (Eds.), Sociolinguistic Perspectives on Register. Oxford University Press, Oxford, pp. 351–385.
Biber, D., Egbert, J., 2023. What is a register?: Accounting for linguistic and situational variation within – and outside of – textual varieties. Register Studies 5 (1), 1–22. https://doi.org/10.1075/rs.00004.bib.
Biber, D., 1988. Variation Across Speech and Writing. Cambridge University Press, Cambridge.
Biber, D., 1994. An analytical framework for register studies. Eds. In: Biber, D., Finegan, E. (Eds.), Sociolinguistic Perspectives on Register. Oxford University Press, Oxford, pp. 31–56.
Biber, D., 1995. Dimensions of Register Variation - A Cross-Linguistic Comparison. Cambridge University Press, Cambridge.
Biber, D., 2012. Register as a predictor of linguistic variation. Corpus Linguistics Linguistic Theory 8 (1), 9–37.
Biber, D., 2019. Text-linguistic approaches to register variation. Register Stud. 1 (1), 42–75.
Cantos Gómez, P., 2013. Statistical Methods in Language and Linguistic Research. Equinox, Sheffield.
CNBC, 2023. Italy Became the First Western Country to Ban ChatGPT. Here's What Other Countries are Doing. from: https://www.cnbc.com/2023/04/04/italy-has-banned-chatgpt-heres-what-other-countries-are-doing.html April 17, 2023.
Conrad, S., Biber, D. (Eds.), 2001. Variation in English: Multi-Dimensional Studies. Longman, Harlow.
Crossley, S., Varner, L.K., McNamara, D., 2014. A multi-dimensional analysis of essay writing: what linguistic features tell us about situational parameters and the effects of language functions on judgments of quality. Eds. In: Berber Sardinha, T., Veirano Pinto, M. (Eds.), Multi-Dimensional Analysis 25 Years on: A Tribute to Douglas Biber. Johns Benjamins, Amsterdam, pp. 344–411.
Davies, M., 2023. Corpus of News on the Web (NOW): 3+ Billion Words from 20 countries, Updated Every Day. https://www.english-corpora.org/now/.
Dutra, D.P., Queiroz, J.M.S., Macedo, L.D.d., Costa, D.D., Mattos, E., 2020. Adjectives as nominal pre-modifiers in chemistry and applied linguistics research articles. Eds. In: Romer, U., Cortes, V., Friginal, E. (Eds.), Advances in Corpus-Based Research on Academic Writing. John Benjamins, Amsterdam.
Forchini, P., 2012. Movie Language Revisited. Evidence from Multi-Dimensional Analysis and Corpora. Peter Lang, Bern.
Friginal, E., Biber, D., 2016. Multi-dimensional analysis. Eds. In: Baker, P., Egbert, J. (Eds.), Triangulating Methodological Approaches in Corpus Linguistic Research. Routledge, Abingdon, pp. 73–89.
Friginal, E., 2013. Twenty-five years of Biber's multi-dimensional analysis: introduction to the special issue and an interview with Douglas Biber. Corpora 8 (2), 137–152.
Goulart, L., Wood, M., 2019. Methodological synthesis of research using multi-dimensional analysis. J. Res. Design Statis. Ling. Commun. Sci. 6 (2), 107–137.
Granger, S., Dupont, M., Meunier, F., Naets, H., Paquot, M., 2020. The International Corpus of Learner English. Version 3. Presses universitaires de Louvain, Louvainla-Neuve.
Gray, B., 2015. Linguistic Variation in Research articles: When discipline Tells Only Part of the Story. John Benjamins, Amsterdam.
Hardy, J.A., 2015. Multi-dimensional analysis of academic discourse. Eds. In: Baker, P., McEnery, T. (Eds.), Corpora and Discourse Studies: Integrating Discourse and Corpora. Palgrave Macmillan, Basingstoke, pp. 155–174.
Hille, P., 2023. AI: Chatbots Replace Journalists. DW. Retrieved 2023-06-21 from. https://www.dw.com/en/ai-chatbots-replace-journalists-in-news-writing/a-65988172.
Köbis, N., Mossink, L.D., 2021. Artificial intelligence versus Maya Angelou: experimental evidence that people cannot differentiate AI-generated from human-written poetry. Comput. Hum. Behav. 114, 106553.

Kublik, S., Saboo, S., 2022. GPT-3: Building innovative NLP Products Using Large Language Models. O'Reilly, Boston.

Le Foll, E, 2021. Register variation in school EFL textbooks. Register Stud. 3 (2), 207–246. https://doi.org/10.1075/rs.20009.lef.

Love, R., Dembry, C., Hardie, A., Brezina, V., McEnery, T., 2017. The spoken BNC2014: designing and building a spoken corpus of everyday conversations. Int. J. Corpus Linguistics 22 (3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov.

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., Liu, X., 2023. *AI* vs. *Human - Differentiation Analysis of Scientific Content Generation*. https://arxiv.org/pdf/2301.10416.pdf.

Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., Zheng, Y., 2019. Recent progress on generative adversarial networks (GANs): a survey. IEEE Access 7, 36322–36333.

Prina Dutra, D., Berber Sardinha, T., 2021. A multi-dimensional typology of English research article sections. Paper presented online at the American Association for Applied Linguistics Conference (AAAL).

Quaglio, P., 2009. *Television Dialogue: The sitcom Friends* vs. *Natural Conversation*. John Benjamins, Amsterdam.

Rascoe, A., Thompson, S., 2023. We Made ChatGPT Write a Song For Us. NPT Retrieved 2023-11-14. from: https://www.npr.org/2023/04/02/1167645526/we-made-chatgpt-write-a-song-for-us.

Theocharopoulos, P.C., Anagnostou, P., Tsoukala, A., Georgakopoulos, S.V., Tasoulis, S. K., Plagianakos, V.P., 2023. Detection of Fake Generated Scientific Abstracts. ArXiv. Issue. https://arxiv.org/abs/2304.06148.

Veirano Pinto, M., & Forchini, P. (to appear). Corpus analysis of the language of film and television. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (2nd ed.). Boston: Wiley.

Veirano Pinto, M., 2013. A linguagem dos filmes norte-americanos ao longo dos anos: uma abordagem multidimensional [The language of American movies over time: A multidimensional analysis]. PhD Dissertation. LAEL, Pontifical Catholic University of Sao Paulo, Sao Paulo.

Veirano Pinto, M., 2019. Using discriminant function analysis in multi-dimensional analysis. In: Berber Sardinha, T., Veirano Pinto, M. (Eds.), Multi-Dimensional Analysis: Research Methods and Current Issues. Bloomsbury, London, pp. 217–230.

Vincent, J., 2023. OpenAI Announces GPT-4 — the Next Generation of Its AI Language Model. Retrieved 2023-10-15. from: https://www.theverge.com/2023/3/14/23638033/openai-gpt-4-chatgpt-multimodal-deep-learning.