Assessment of different GPT Models versus the human text: a quantitative analysis of lexis and cohesion

Ocena różnych modeli GPT względem tekstu napisanego przez człowieka: analiza leksyki i spójności tekstu

Dominik SKOWRON[1]
University of Economics in Krakow

Anna BĄCZKOWSKA[2]
University of Gdańsk

Abstract

The aim of the paper is to analyse select lexical parameters and text cohesion of academic texts automatically generated by several OpenAI LLM models (GPT) in order to (1) investigate the quality of GPT output relative to the original text; (2) compare the quality of various GPT models. The material used in the study comprised a fragment of a research article. The method of analysis involved NLP-based text analysis tools that focus on the examination of various lexical and text cohesion parameters. The study shows that the similarity between AI-generated texts and the text written by the human author is very high and that there is no single model which would achieve the highest values for all linguistic indices.

Keywords: GPT, LLM, ChatGPT, OpenAI, cohesion, lexical variation

Streszczenie

Celem artykułu jest automatyczna analiza wybranych aspektów leksykalnych oraz kohezji w tekstach akademickich wygenerowanych przez kilka modeli OpenAI LLM (GPT) w celu (1) zbadania jakości tekstów sparafrazowanych przez GPT w porównaniu do tekstu oryginalnego; (2) porównania jakości różnych modeli GPT względem siebie. Materiał użyty w niniejszym badaniu zawiera fragment tekstu akademickiego. Metoda użyta w badaniu wykorzystuje narzędzia analizy tekstu oparte na przetwarzaniu języka naturalnego, które skupiają się na porównywaniu różnych parametrów leksykalnych i spójności tekstu. Badanie pokazuje, że podobieństwo między tekstami wygenerowanymi przez sztuczną inteligencję a tekstem oryginalnym jest bardzo wysokie oraz że nie istnieje pojedynczy model, który osiągnąłby najwyższe wartości we wszystkich parametrach.

Słowa kluczowe: GPT, LLM, ChatGPT, Open AI, spójność tekstu, różnorodność leksykalna

[1] University of Economics in Krakow, Poland
https://orcid.org/0009-0003-0143-3457
dominikskowron007@gmail.com

[2] University of Gdańsk, Institute of English and American Studies, Poland
https://orcid.org/0000-0002-0147-2718
anna.baczkowska@ug.edu.pl

1. Introduction

The aim of the study is to verify the quality of texts automatically generated by several GPT models relative to the original text on the one hand and to compare the differences among the models on the other. The texts will be examined against select parameters of lexical features and text cohesion. The AI tool (OpenAI LLM GPT Models) and its types and models will be presented in section 2, and the parameters employed in this study to scrutinise the data selected for the experiment will be elaborated in section 3. The research aims and methodology employed here will be described in section 4. The results will be delineated in section 5, and they will demonstrate distinctive features of all examined texts in line with 240 indices that comprise lexical and discourse aspects of the texts at issue.

2. Open AI LLM Models (GPT)

Large Language Models (LLMs), such as the Generative Pre-trained Transformer (GPT) models, have emerged as a significant breakthrough in the field of Natural Language Processing (NLP). These LLMs are designed to generate coherent and natural-sounding text that is highly similar to human-generated text. The GPT models, in particular, are known for their ability to create high-quality text in a wide range of styles and domains due to their transformer-based neural network architecture, which predicts the next word or sequence of words based on the input text context.

The pre-training of these models is accomplished through exposure to massive amounts of textual data, which allows the models to learn and capture the underlying patterns and structures of natural language. Once trained, the models can be fine-tuned for specific applications and domains, such as language translation, chatbots, text summarisation, and content generation, among others.

Due to their advanced capabilities, LLMs have the potential to transform the landscape of NLP and significantly impact various industries and sectors. This article aims to provide a brief overview of the GPT models (Table 1), including their architecture, pre-training process, and applications, as well as their limitations and potential areas for future research.

We have selected five models enumerated below for our investigation. Table 1 shows their basic features and how they differ.

Model name	Model	Description	Training data
text-Ada-001	GPT-3	Capable of very simple tasks, usually the fastest model in the GPT-3 series and lowest cost	Up to Oct 2019
text-Babbage-001	GPT-3	Capable of straightforward tasks, very fast, and lower cost.	Up to Oct 2019
text-Curie-001	GPT-3	Very capable, faster and lower cost than Davinci.	Up to Oct 2019
text-Davinci-003	GPT-3.5	Can do any language task with better quality, longer output, and consistent instruction-following than the Curie, Babbage, or Ada models. It also supports inserting completions within text.	Up to Jun 2021
ChatGPT	GPT-3.5	GPT-3.5 ChatGPT: A highly advanced language model based on the GPT-3.5 architecture, excelling at a wide variety of language tasks with improved quality and consistency over its predecessors. Its robust capabilities include generating meaningful responses, engaging in complex conversations, and providing concise summaries, making it suitable for diverse applications.	Up to 2021; it can continue to learn and improve over time as it interacts with users and is fine-tuned with new data.

Table 1. AI models

3. Lexical parameters and text cohesion

Lexical parameters of a text engulf a number of potential aspects that measure, inter alia, lexical density and diversity. A high result achieved within these

indices speaks for a good quality text, a varied vocabulary and advanced lexical knowledge. In this study, lexical variation (LV) and lexical density (LD) will be analysed with the help of Natural Language Processing tools, which automatically extract such data.

LV assesses how varied words are used in a text by dividing the number of unique words (the so-called types) by the number of all words (i.e., tokens); this value also known as TTR (type-token ratio) or lexical diversity. The problem with TTR measure resides in the fact that it is text length sensitive, wherein the longer the text, the lower values are produced. Hence, a number of other formulas have been offered, e.g., the index of Guiraud (1960), which involves taking the square root of TTR, vocd-D (Malvern and Richards, 1997), the Maas index (Mass, 1972), or MTLD (McCarthy, 2005). More stable and reliable results, however, are achieved with MATTR and HD-D (Zenker and Kyle, 2021). MATTRx relies on moving the average type-token ratio in an x-word window, while HD-Dx resorts to the hypergeometric distribution in order to calculate the probability of finding a token in a random sample of x tokens. LD in turn, takes into account the number of content words against function words; the more content words, the more lexically dense the text. In other words, LD manifests lexical richness/complexity (Zenker and Kyle, 2021).

Text cohesion refers to explicit textual clues that connect different segments of text and ideas expressed therein. These may include overlapping sentences or paragraphs (and may span two or more sentences/paragraphs), connectives (which link ideas), semantic similarity between paragraphs and other cohesive devices. They inform readers that similar ideas are communicated across consecutive sentences and paragraphs. The overall text similarity is measured by resorting to the parameter of keyness or by extracting information by means of advanced AI-based calculations (such as word2vec). Along with these, several other indices were considered in the present study. Firstly, Latent Semantic Analysis (LSA; Landauer et al., 1998), which is based on cosine similarity and measures semantic overlap between sentences or paragraphs by analysing "explicit words and words that are implicitly similar or related in meaning" (McNamara et al., 2014: 66). Secondly, Latent Dirichlet Allocation divergence score (LDA; Blei et al., 2003), which is based on topic modelling. Keyness is another parameter that describes text cohesion. Keyness offered by TAACO relies on COCA (Davis, 2008) data gleaned from magazines and news sections. TAACO also analyses the so-called

givenness of information. It has been noticed that cohesive devices encoding givenness are reliable predictors of good-quality essays and general text cohesion (Crossley et al., 2016), as well as higher text comprehensibility inasmuch as given information refers back to what was previously mentioned in a text (Crossley et al., 2014: 518). Givenness entails the use of such words as determiners, pronouns and demonstratives (Crossley et al., 2014), which replace nouns, i.e., they are words which make information recoverable from preceding discourse. In addition to that, a high value for noun-pronoun ratio is indicative of high-quality texts. Based on WordNet database, semantic overlap in turn calculates the overlap between words and word synonyms at sentence and paragraph levels. Cohesive devices such as semantic overlap and the use of connectives (e.g., conjunctions, disjunctions, sentence linking phrases/words) are also associated with general high-quality writing (McNamara et al., 2010; McNamara et al., 2014: 108), while measures of paragraph-to-paragraph semantic overlap are positively correlated with text coherence (McNamara et al., 2014: 112).

It must be noted that the term cohesion is often juxtaposed (or mistaken) with the concept of coherence. Contrary to cohesion, coherence focuses on the text as a whole, on how it can be understood by a reader. Coherence involves textual and extratextual clues, such as the reader's prior knowledge and reading skills. This study aims to examine cohesion only.

4. Aim, material and methods

4.1. Aim

The aim of the study is to check how artificial intelligence copes with producing a text based on its source, which is an academic text written by a human author. The source text is a fragment extracted from an original research paper randomly selected from a high-quality journal specialising in comparative pragmatics. As the AI models are believed to produce texts of a high degree of similarity to human-generated high-quality texts, we wanted to verify this general public opinion. The null hypothesis that there are no differences between the source text and the GPT-based texts will be verified by analysing specific features at the lexical and text-cohesion levels within select parameters. Along with comparing AI-generated texts with the source text, the second aim is to unveil the possible differences in the quality of texts produced by specific GPT models. It is assumed that the early models (Ada,

Babbage) will not cope with text generation to the same high-quality degree as the more recent ones, i.e., Davinci and in particular ChatGPT.

4.2. Material

The first step in this study involved identifying the appropriate article for analysis. We chose a scientific article for our investigation, as our aim was to verify academic texts. Specifically, we selected a fragment of a paper published in a linguistic journal (*Contrastive Pragmatics*) titled "Variation Patterns in Interlanguage Pragmatics: Apology Speech Act of EFL Learners vs. American Native Speakers" (Eslami et al., 2022). The selected fragments included the Introduction, Literature Review, Discussion, and Conclusion. The analytical part was skipped as we wanted to retrieve longer, original running text undisturbed by exemplifications and statistical analysis.

Subsequently, OpenAI GPT models were utilised to rephrase the identified texts. To fit within the token restrictions of the models, which allow 2048 tokens per input, the texts were chunked into smaller segments. Uniform chunks were used across all models, and the OpenAI Playground option was chosen due to the relatively small amount of data involved. While an API by request was also available, it was deemed unnecessary.

The most critical consideration during the text rephrasing process was the choice of prompt. Rather than constructing the most effective prompt, we opted for a simple prompt to obtain the default and the most likely to be common answers. The prompt comprised two elements: "Summarize: " and a chunk of text. Temperature, top P and other parameters set to default. For each of the GPT models, the text chunk was inputted, and outputs were generated.

One of the challenges encountered during the process was the occasional production of outputs that were either too short or did not align with other models. To overcome this issue, we continued to ask the model to provide outputs until satisfactory results were obtained. Consequently, we were able to generate text files from the selected article for every researched OpenAI Large Language model.

4.3. Methods

The data were analysed with the aid of two NLP tools designed for quantitative linguistic analysis, namely TAACO (Crossley et al., 2016) and TAALED (Kyle and Crossley, 2015). These NLP-based software tools for language analysis contain a high number of specific indices which allow one to delve into the

detailed linguistic categories at word, sentence, paragraph and text levels. NLP tools extract and measure linguistic features automatically; thus, they replace time-consuming and labour-intensive manual tagging of, e.g., word classes, syntactic structures, textual aspects, and readability parameters. TAACO measures 201 indices of text cohesion relying on lexical and semantic overlaps, text similarity, givenness and connectives, with some basic lexical parameters included. TAALED in turn focuses on 39 lexical parameters. Altogether, 240 indices were used to compare the human-generated text with the five AI-generated models of writing based on the human text.

5. Research results

5.1. Lexical parameters

The general observation of lexical parameters starts with the number of words in each text; specifically, we analysed the number of all words (tokens), all types (unique words), content types and tokens and function types and tokens (Fig. 1). Interestingly, the most recent AI model, ChatGPT, substantially shortened the original text, while Babbage is closest to the source text in terms of the number of various types and tokens.

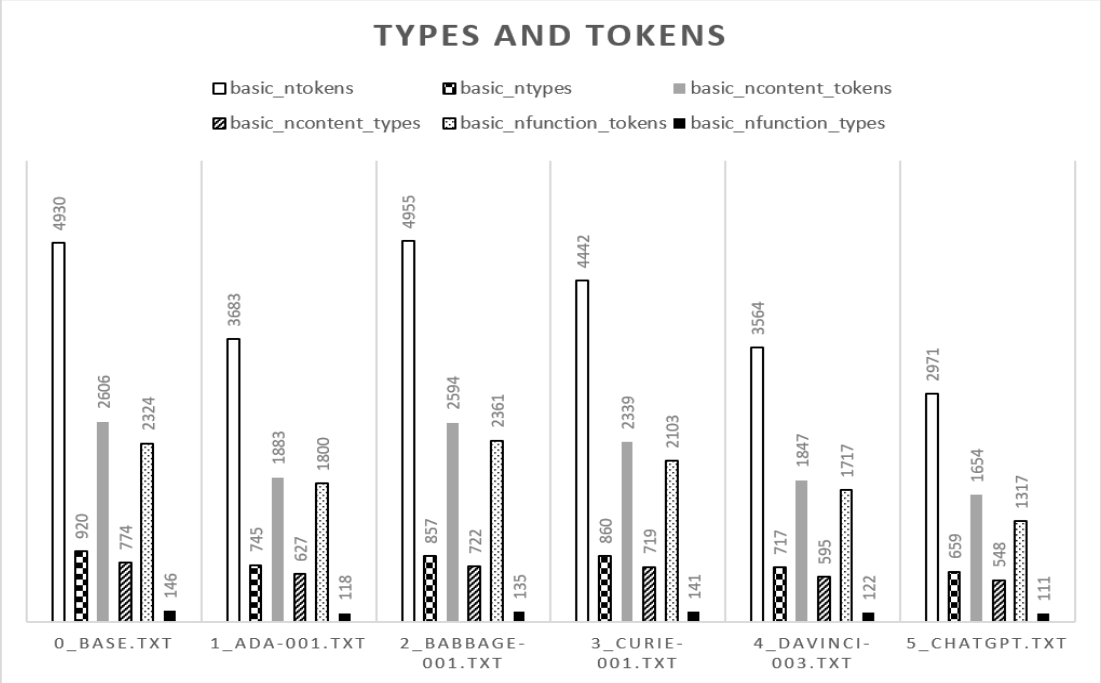

Fig 1. Types and tokens across all texts for all words, content words and function words

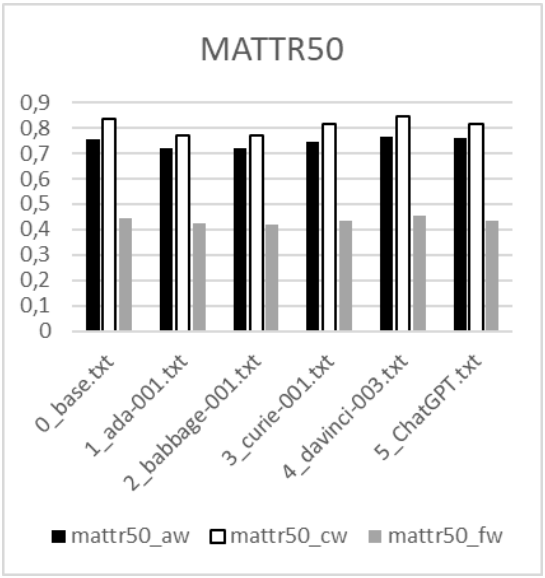

Fig 2. MATTR50 for all words, content words and function words

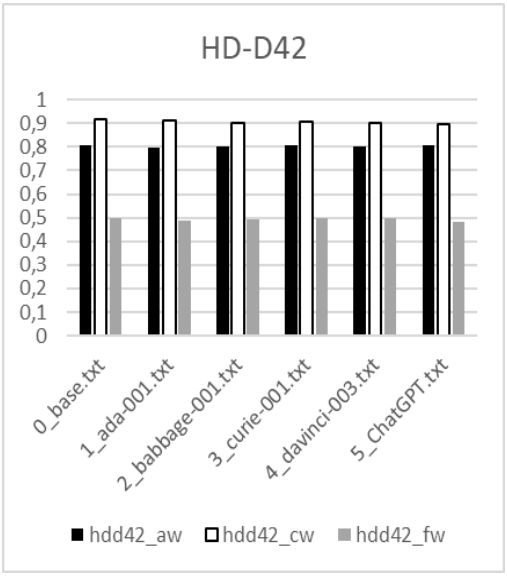

Fig 3. HD-D42 for all words, content words and function words

Even a cursory examination of the outcomes allows noticing no substantial differences with regard to the source text within the categories of content words, function words and all words (with only minor divergence for content words in MATTR50 outcome). This intuitive judgment was confirmed by Kruskal-Wallis test (H = 0.2057, p = .999, $\eta 2$ = -.073), with a very small effect size, and high chances of type I error in case of rejecting H_0. This means that lexical variations of all GPT texts show little differences and that the GPT texts bear a high similarity to the human-generated text. It was thus safe to apply further TTR-based calculations, such as the TTR of unique content word lemmas, function word lemmas, as well as pronoun lemmas, noun lemmas, verb lemmas, adjective lemmas and adverb lemmas (Fig. 4), including n-gram lemmas (bigrams and trigrams), in order to observe more subtle discrepancies.

The results for the categories illustrated by Fig. 4 display a minor leap of adverb lemmas TTR for the Curie model and preposition lemmas TTR for Davinci and ChatGPT. Lexical density of both types and tokens is also very similar across all versions.

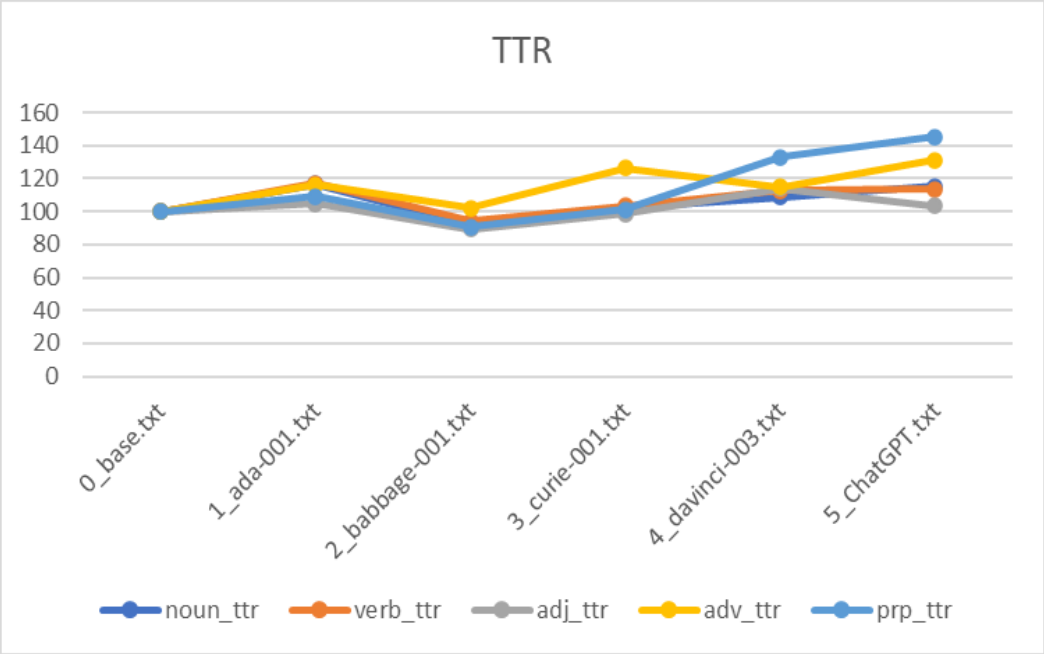

Fig 4. TTR for nouns, verbs, adjectives, adverbs and pronouns in percentage

Regarding the number of tokens, the best result for these parameters is achieved by Babbage, which used almost the same number of words, and the greatest dissimilarity can be observed in the case of ChatGPT, which significantly reduced the length of the paraphrase relative to the original text. This reduction stems from the prompt used to generate the AI texts ("summarise") to which ChatGPT proved most responsive.

5.2. Cohesion

First and foremost, it must be noted that while some differences among the texts have been noticed, they are not statistically significant and thus show only some tendencies. In other words, all the texts generated by AI show a high similarity to the source, human-generated text. Some minor discrepancies that signal certain tendencies, which should receive further validation based on different texts, are discussed below.

5.2.1. Text similarity

All three indices offered by TAACO have been used to check the overall text similarity, i.e., LSA, LDA, and word2vec. Regardless of which index is used, all

AI-generated texts display an extremely high resemblance to the original human-generated source text, which oscillates around 100%. The highest scores in AI-generated texts are earned by Curie and the lowest by Ada (ca. 98% for LSA).

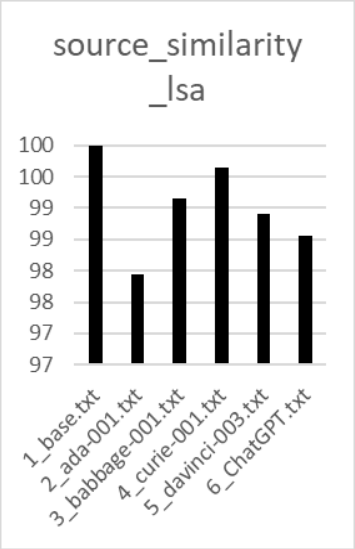

Fig 5. Overall text similarity - LSA in percentage

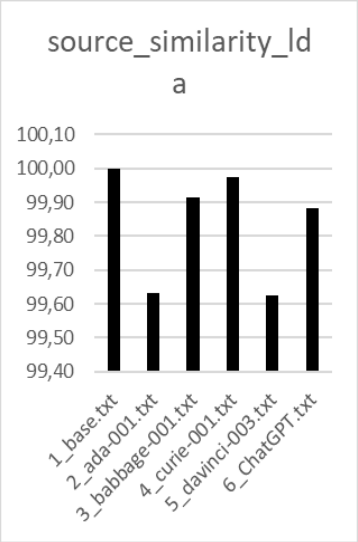

Fig 6. Overall text similarity - LDA in percentage

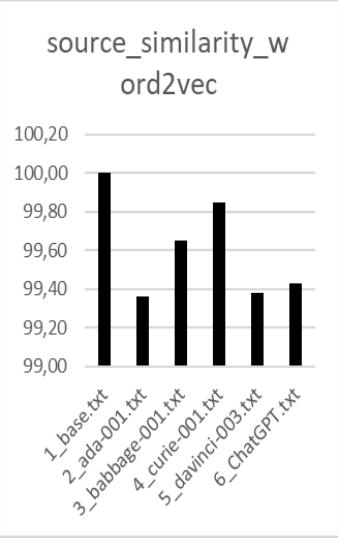

Fig 7. Overall text similarity - word2vec in percentage

Measured by means of keyness, text similarity earned similar scores. Keyness was measured for nouns, verbs, and adjectives, for unigrams, bigrams, trigrams and quadgrams. There is a great internal consistency of the data generated by AI, yet Ada stands out by generating a text of the least resemblance to the original text. For unigrams (with Cronbach's $\alpha = .897$), there is a high similarity among the models, as well as for bigrams ($\alpha = .926$), trigrams ($\alpha = .913$) and quadgrams ($\alpha = .935$).

The similarity of keywords among various AI versions thus holds for four types of n-grams, i.e., it spans words and word combinations. From this, it transpires that, thematically, the texts strongly resemble the original text due to the usage of words/phrases typical of the topic elaborated by the human author.

5.2.2. Givenness

Givenness was verified by means of the pronoun-noun ratio (PNR) index, which divides the number of third-person pronouns by the number of nouns.

A high PNR score indicates high text comprehensibility (i.e., ease of text processing and understanding), as given information is easier to grasp than new information. In the texts under inspection, PNR displays the most significant divergence from the source text in the case of Ada (136%), which made the text the easiest to follow. ChatGPT (73%) also drifts away from the base text, yet the score drops relative to the original text, which is suggestive of lowering the comprehensibility of the text. Although in TAACO givenness does not comprise determiners and demonstratives, they are treated as indicators of givenness by Crossley et al. (2015), and hence were taken into consideration here. Both indices proved the closest to the source text in the case of ChatGPT. Other AI models scored higher for the remaining givenness indices than the human-generated text, which shows that they try to make the originally rather advanced texts (academic writing, scientific journal) more accessible, i.e., easier to read.

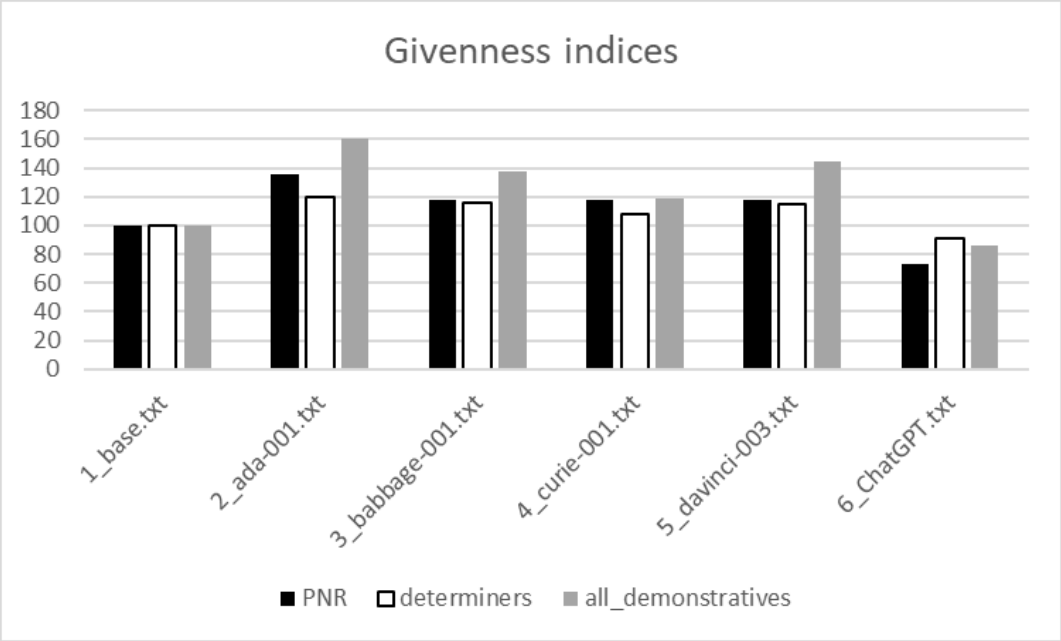

Fig 8. Givenness indices in percentage: pronoun-noun ratio, determiners and demonstratives in percentage

5.2.3. Connectiveness

Lexical subordination is typically realised by words such as *although*, *as*, which initiate a clause that is subordinate to another clause. Interestingly, of all AI

texts analysed here, only ChatGPT scored slightly lower than the source text (92%), while other models scored higher values, with Ada reaching the highest (135%) score. Sentence linking (e.g., *nonetheless*, *therefore*) exceeds the original text yet to a small degree, and order words (e.g., *to begin with*, *next*, *first*) reached higher values for all AI-texts, with Ada obtaining 262% of the original text. Similarly, temporal connectives (*a consequence of*, *again*, *after*) earned scores higher for AI, with the highest score of 199% of the original text for Davinci. On the other hand, the original text has more coordinating conjunctions than all AI texts. Ada, again, stands out the most, with the 61% of the score for the original text. Logical connectives (*actually, admittedly*, *after all*) have lower scores than the original text for all GPT models (Babbage – 96%, Curie – 94%, Davinci – 84%, ChatGPT – 76%), except for Ada (103%).

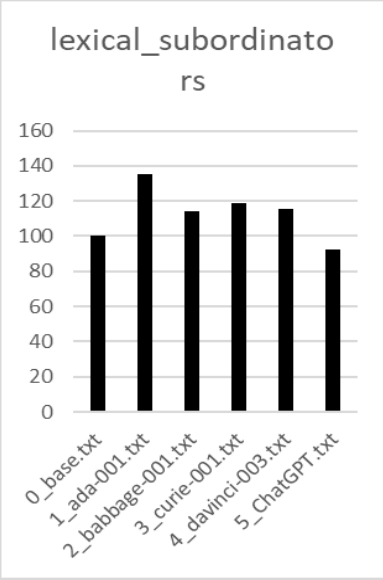

Fig 9. Lexical subordinators to all words ratio in percentage

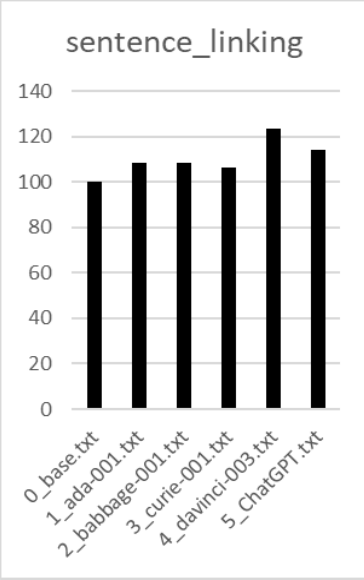

Fig 10. Sentence linking words to all words ratio in percentage

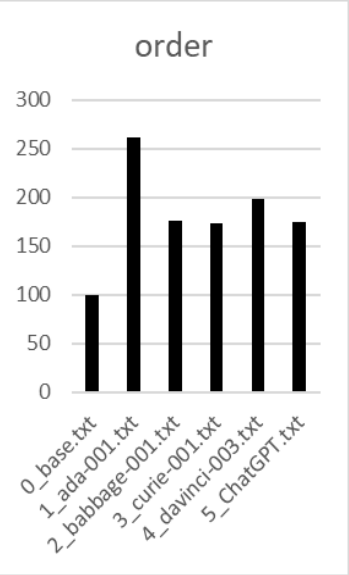

Fig 11. Order words to all words ratio in percentage

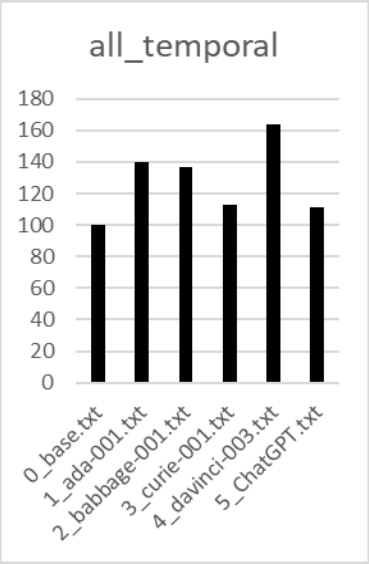

Fig 12. Temporal connectives to all words ratio in percentage

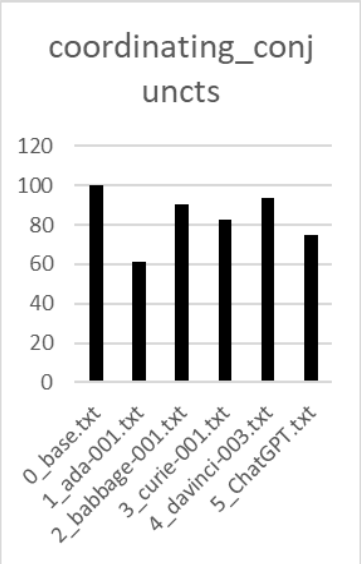

Fig 13. Coordinating conjuncts to all words ratio in percentage

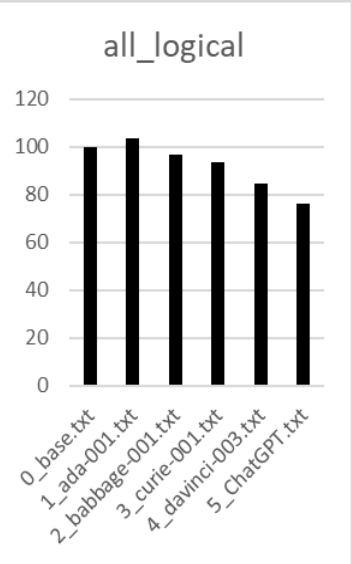

Fig 14. All logical connectives to all words ratio in percentage

Apart from the select connectives presented above, all the connectives offered by TAACO (25 indices) have been analysed as aggregated results. In consequence, it can be observed that the earlier AI models (Ada, Babbage and Curie) diverge from the source in terms of all connectives taken together inasmuch as they tend to use them more seldom than the human writer. The more recent model Davinci gets very close to the source, while ChatGPT exceeds the original text in the use of connectives.

5.2.4. Semantic overlap

Sentence-to-sentence overlap of synonymous nouns tends to be higher for GPT than for the original, whereas of synonymous verbs tends to be lower. The lowest verb overlap is in the case of Ada (40% relative to the original text), and the highest noun overlap is taken by ChatGPT (160%). Some AI texts (Ada, Babbage and ChatGPT) resort to more synonymous nouns used between adjacent sentences, while all AI texts employ a smaller number of synonymous verbs. ChatGPT is most varied in the choice of nouns measured sentence-to-sentence.

As for paragraph-to-paragraph overlap, the value is smaller for nouns and verbs. Inter-paragraph cohesion is thus lower for IA texts than the human text. Overall, however, when average values for all words are considered, the outcomes for LSA, LDA and word2vec revolve around 100%, which speaks for a generally very high score earned for semantic overlap. In other words, all AI texts display high-quality writing features measured by means of semantic overlap.

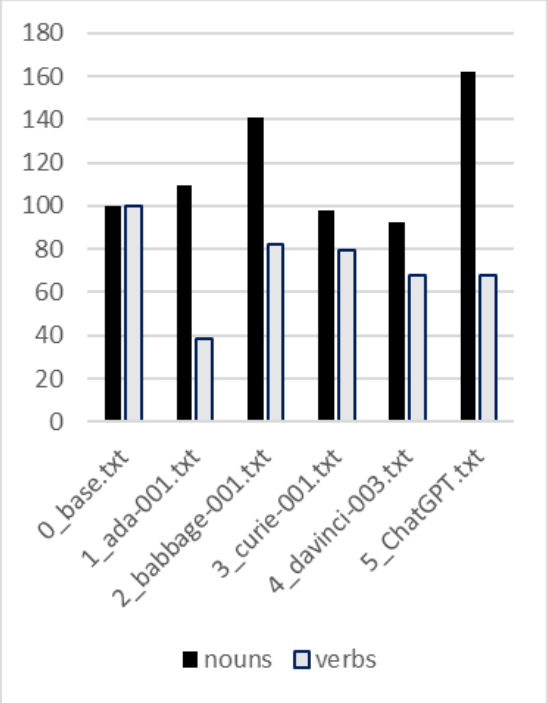

Fig 15. Average sentence-to-sentence overlap of noun synonyms and verb synonyms in percentage

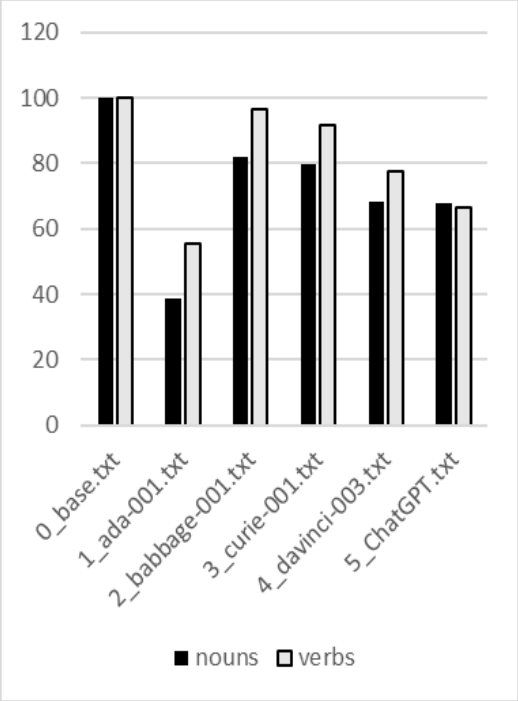

Fig 16. Average paragraph-to-paragraph overlap of noun and verb synonyms in percentage

In sum, cohesion based on semantic overlap is mainly achieved at the inter-sentential level. The meaning overlap between adjacent paragraphs does not seem to be controlled in AI texts.

5.2.5. Lexical overlap at sentence and paragraph levels

One of the variables that measure lexical overlap at the sentence level is adjacent sentence overlap for lemmas. It divides the number of lemma types that occur at least once in the next sentence by the number of types in each sentence. This parameter is higher in the case of all AI texts relative to the

source text (Fig. 17), with the highest score for ChatGPT (129%) and Curie (124%) and the lowest for Davinci (104%). Similarly, the same parameter with the exclusion of the last sentence, i.e., sentence-normed, reaches values higher than the original text: ChatGPT scored 137%, Babbage 116%, while Davinci 101%. ChatGPT seems to care about maintaining inter-sentential cohesion (for single-incidence words) by repeating unique words, even more than the human author. A reverse result is discernable for the span of three sentences for any lemma (whether types or tokens). Here, the original text gets the highest score and ChatGPT the lowest (Fig. 18). From this, it transpires that AI texts tend to repeat unique words across sentences (which boosts the ease of reading) and avoid using the same words (which enhances text quality). In this respect, it seems that AI may gain even better results than human text. Judging by the results presented in Fig. 8, these unique word types can comprise pronouns, determiners and demonstratives, yet other results demonstrate that noun, verb (except for Davinci), adjective (except for Davinci) and adnoun (except for Davinci) types as well as noun synonyms (see Fig. 15) are also repeated across sentences by AI. Particularly high values are obtained for ChatGPT for the noun and adjective types (176% and 179% respectively).

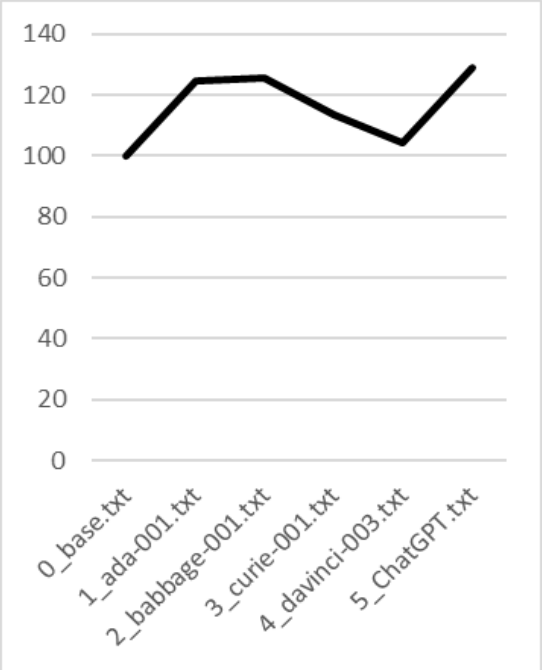

Fig 17. Overlap of lemma types in two adjacent sentences

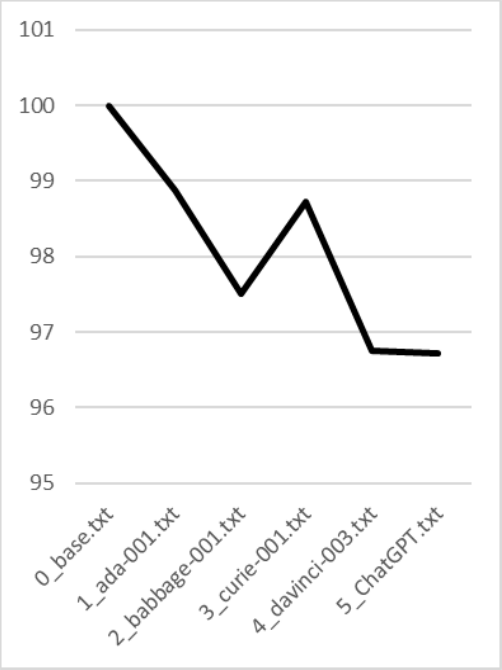

Fig 18. Overlap of all lemmas in three adjacent sentences

Regarding lexical overlap at the paragraph level, the most conspicuous values diverging from the source texts (five indices) are presented in Fig. 19. As can be seen, all indices reached values below the 100% represented by the original text, except for one related to the incidence of pronoun lemmas (paragraph normed). The pronoun index measures the number of pronoun lemma types relative to the number of paragraphs. Put simply, the number of unique pronouns occurring at least once in the subsequent paragraph is higher in AI-driven texts than in the penned writing, except for ChatGPT. This is suggestive of text ease enhancement in the case of AI output since, as already mentioned, by deploying pronouns a text is more comprehensible and easier to read. On the other hand, the overlap of unique adjectives and adnouns across two paragraphs as well as unique nouns and verbs across three paragraphs is lower in the case of AI. Accordingly, in AI-driven texts, words are not repeated across paragraphs.

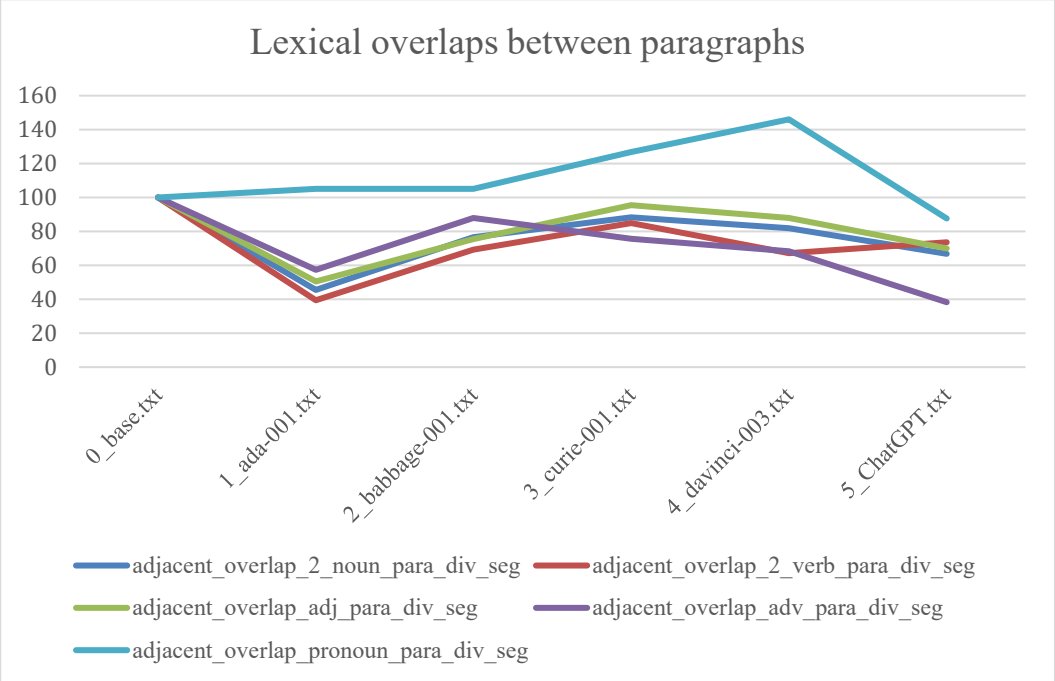

Fig. 19. Paragraph normed lexical overlaps between two or three paragraphs in percentage

To sum up this section, based on the single text at issue, it seems that the text written by a human author tends to repeat word tokens in subsequent sentences, while AI appears to have a tendency to abstain from deploying the

same words and to refer to words that occurred only once in the previous sentence. The human writer repeated unique words at the paragraph level while AI at the inter-sentential level.

5.2.6. Overall analysis

Of all the 240 indices used to analyse various lexical and textual features, 60 were selected for the overall analysis. The indices that were left out showed minute diverges relative to the original text; hence, they were ignored. The 60 selected indices manifest differences which are not statistically significant yet display some visible discrepancies, i.e., tendencies that could give a hint for further investigation in some future studies. Kruskal-Wallis test for non-parametric data was applied to the results obtained for all six texts in order to observe whether any of the indices drifted away from the other scores. The result (H = .4692, p = .9932, $\eta 2$ = - 0.0063) testifies that the chance of type I error is very low, and thus we can safely accept the H_0. By the same token, the study confirms the expected outcome that AI-generated texts bear a high resemblance to the text penned by the human author.

6. Conclusion

The study has shown that, firstly, there are no statistically significant differences between the original text written by the human writer and any of the texts automatically generated by five models of AI. The results thus prove that the effectiveness of AI in imitating the human author is very high. Secondly, regarding cohesion, ChatGPT seems to be the closest to the original text, while within lexical parameters, Davinci scored best. Overall, the text created by ChatGPT resembles the original text the most, and Ada, diverges from other models, often obtaining scores opposite to either other AI models or the original text.

Furthermore, in some aspects, ChatGPT seems to show a tendency to achieve very good results, even better than the original text; these refer to, e.g., limiting repetitions of words occurring several times across adjacent sentences (i.e., tokens) and repeating newly-appeared words (i.e., types) in subsequent sentences (or substituting them with synonyms in the case of nouns). Both tendencies may contribute to ease of reading. A higher text comprehensibility and cohesion can be also achieved by deploying pronouns replacing nouns and by utilizing demonstratives and determiners (dubbed givenness of information). In this respect, GPT models display higher values

than the human author, except for ChatGPT, which more often resorted to nouns, verbs and adjectives. Moreover, sentence linking words, order words and temporal connectives are more abundant in the AI-driven texts, and all types of connectives taken together are more often used by the two recent AI models (Davinci and ChatGPT).

This preliminary study has shown some interesting tendencies in the differences across AI texts relative to the original one, yet it has some limitations, namely, the fact that only one original text was taken into consideration and that its length was limited. As a result, the current study needs further validation on more and longer texts. Nonetheless, the outcome of this investigation could guide future studies as it already highlights specific indices where the scores diverged the most from the source text.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods 48*(4), 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

Davies, M. (2008). The corpus of contemporary American English. Provo, UT: Brigham Young University.

Eslami, Z. R., Raeisi-Vanani, A., and Anani Sarab, M. R. (2022). Variation Patterns in Interlanguage Pragmatics: Apology Speech Act of EFL Learners vs. American Native Speakers. *Contrastive Pragmatics*, *4*(1), 27-63. https://doi.org/10.1163/26660393-bja10068

Guiraud, P. L. (1960). Problèmes Et Méthodes De La Statistique Linguistique/ Problems and Methods of Statistical Linguistics. Dordrecht: Kluwer.

Kyle, K. and Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly 49*(4), 757-786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., and Jarvis, S. (2021). Assessing the validity of lexical diversity using direct judgements. *Language Assessment Quarterly 18*(2), 154-170. https://doi.org/10.1080/15434303.2020.1844205

Maas, H.D. (1972). Zusammenhang zwischen Wortschatzumfang und Langeeines Textes. *Zeitschnftfur Literaturwtssenschaft und Linguistik*, 8, 73-79.

Malvern, D. D. and Richards, B. J. (1997). A new measure of lexical diversity. In Ryan, A. and Wray, A. (Eds), *Evolving Models of Language*. Clevedon: Multi-lingual Matters, 58-71.

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity* (Doctoral dissertation). Available from Proquest Dissertations and Theses. (UMI No. 3199485)

McNamara, D. S., Crossley, S. A., and McCarthy, P. M. (2010). The linguistic features of quality writing. Written Communication, 27, 57–86.

McNamara, D., S., Graesser, A., C., McCarthy, P. M., and Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284. https://doi.org/10.1080/01638539809545028

McNamara, D. S., Crossley, S. A., McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication* 27, 57-86.

Zenker, F., Kyle, K**.** (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing 47*. https://doi.org/10.1016/j.asw.2020.100505