RAEL Revista Electrónica de Lingüística Aplicada

RAEL: Revista Electrónica de Lingüística Aplicada
Vol./Núm.: 23/1
Enero-diciembre 2024
Páginas: 34-54
Artículo recibido: 14/08/2024
Artículo aceptado: 01/12/2024
Artículo publicado: 31/01/2025
Url: https://rael.aesla.org.es/index.php/RAEL/article/view/666
DOI: https://doi.org/10.58859/rael.v23i1.666

¿Tienen GPT-3.5 y GPT-4 un estilo de escritura diferente del estilo humano? Un estudio exploratorio para el español

Do GPT-3.5 and GPT-4 Have a Writing Style Different from Human Style? An Exploratory Study for Spanish

Lara Alonso Simón
Ana María Fernández-Pampillón Cesteros
Marianela Fernández Trinidad
Manuel Márquez Cruz
Universidad Complutense de Madrid

La cuestión que se aborda en este trabajo de investigación es la comprobación, mediante técnicas estadísticas, de que los modelos generativos de lenguaje GPT-3.5 (versión gratuita) y GPT-4 (versión de pago) de ChatGPT tienen un estilo de escritura distinto al de los humanos, y que pueden diferenciarse, al menos, por tres tipos de rasgos: léxicos, signos de puntuación y estructura sintáctica de las oraciones. Determinar si los grandes modelos de lenguaje tienen un estilo propio es relevante de cara a poder detectar la autoría automática de los textos. En trabajos anteriores se construyó un corpus comparable de textos humanos y automáticos en español y, mediante un estudio cualitativo, se localizó un conjunto de rasgos lingüísticos y estilísticos propios de cada autor. En este trabajo se ha podido comprobar cuantitativamente que 17 variables lingüísticas presentan diferencias estadísticamente significativas entre autores humanos y los modelos GPT-3.5 y GPT-4.

Palabras clave: *estilo de escritura; grandes modelos de lenguaje; GPT-3.5; GPT-4; lingüística de corpus*

The aim of this research is to verify, using statistical methods, that the generative language models GPT-3.5 (free version) and GPT-4 (paid version) of ChatGPT have their own writing style distinct from that of humans and that they can be distinguished by at least three types of features: lexical features, punctuation marks and syntactic sentence structure. Determining whether large language models have their own style is relevant in order to detect automatic authorship of texts. In previous work, a comparable corpus of human and automatic texts in Spanish was constructed and, through a qualitative study, a set of linguistic and stylistic features specific to each author was identified. In this work, it has been quantitatively demonstrated that 17 identified linguistic variables show statistically significant differences between human authors and the GPT-3.5 and GPT-4 models.

Keywords: *writing style; large language models; GPT-3.5; GPT-4; corpus linguistics*

Citar como: Alonso Simón, L., Fernández-Pampillón Cesteros, A.M., Fernández Trinidad, M. y Márquez Cruz, M. (2024). ¿Tienen GPT-3.5 y GPT-4 un estilo de escritura diferente del estilo humano? Un estudio exploratorio para el español. *RAEL: Revista Electrónica de Lingüística Aplicada, 23*, 34-54. https://doi.org/10.58859/rael.v23i1.666

1. INTRODUCCIÓN

Los grandes modelos de lenguaje (en inglés LLM – *Large Language Models*) son modelos estadísticos de lenguaje creados mediante el entrenamiento de grandes redes neuronales, con enormes cantidades de texto, y utilizando potentes recursos computacionales (Hadi, Al-Tashi, Qureshi, Shah, Muneer, Irfan, Zafar, Shaikh, Akhtar, Wu y Mirjalili, 2023). Actualmente, estas redes neuronales de varias decenas o centenas de miles de conexiones están creadas utilizando una arquitectura (o variaciones de ésta) denominada Transformador (en inglés *Transformer*) diseñada por Google (Vaswani, Brain, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser y Polosukhin,, 2017). Cada una de estas conexiones tiene un peso, representado por un valor numérico de tipo real, que se calcula mediante el entrenamiento. El peso de cada conexión indica la importancia de esa conexión para entender o producir algún componente del lenguaje natural. Así, un gran modelo del lenguaje no es más que un conjunto de cientos de miles de parámetros numéricos aprendidos durante el entrenamiento que contienen implícitamente los patrones y las estructuras lingüísticas de los textos del entrenamiento.

Si bien un mismo LLM puede aplicarse a múltiples tareas de Procesamiento del Lenguaje Natural (en adelante PLN) como la clasificación, la generación de texto o la traducción automática, podemos distinguir entre los LLM creados para entender el lenguaje y aplicados, preferentemente, a tareas de clasificación –p. ej. el modelo del español BETO (Cañete, Chaperon, Fuentes, Ho, Kang y Pérez, 2020)– y los LLM creados para generar lenguaje –p. ej. los modelos de la familia GPT (Radford, Narasimhan, Salimans y Sutskever, 2018)–. En este trabajo nos centramos en los LLM generativos y, en particular, en dos modelos de la familia GPT, GPT-3.5 y GPT-4. Estos modelos subyacentes al sistema ChatGPT pueden generar texto multilingüe y, en el caso del español, con una calidad muy alta.

Los LLM generativos son capaces de producir textos tan coherentes y contextualmente relevantes que es difícil distinguirlos de los textos escritos por humanos (Uchendu, Le y Lee, 2023). De hecho, la investigación de Casal y Kessler (2023) revela que la tasa global de identificación positiva para distinguir textos generados por ChatGPT (GPT-4 Plus versión de pago) de los textos generados por humanos fue solo del 38,9 % en una evaluación realizada por 72 revisores de las 30 principales revistas de Lingüística Aplicada.[1] Así, la alta calidad del lenguaje que generan ha supuesto una mejora sin precedentes de las aplicaciones de PLN (Hadi et al., 2023).

Ocurre, sin embargo, que los LLM generativos también se utilizan con fines maliciosos, como la generación de noticias falsas (desinformación) y reseñas sobre productos y servicios (desprestigio), el plagio en el ámbito académico (suplantación) o el envío de correos electrónicos de amenaza o engaño (Pizarro, 2019; Pavlyshenko, 2022; Cardenuto, Yang, Padilha, Wan, Moreira, Li, Wang, Andaló, Marcel y Rocha, 2023; Crothers, Japkowicz y Viktor, 2023). Es, en estos casos, cuando resulta necesario encontrar métodos y herramientas que ayuden a detectar el uso malicioso de la generación automática de textos (Maloyan, Nutfullin y Ilyushin, 2022). La detección automática o semiautomática de estos textos generados por LLM es un problema difícil, y aún no resuelto, que tiene gran impacto en todos los sectores: político, económico, académico y social (Uchendu et al., 2023).

Nuestro objetivo es contribuir a mejorar la identificación de los textos generados automáticamente. La hipótesis de partida se basa en considerar que cada LLM tiene su propio estilo de escritura y que existen diferencias significativas en cómo utilizan el lenguaje las personas y los LLM. En este trabajo exploramos la viabilidad de la hipótesis evaluando dos LLM utilizados por ChatGPT para generar texto en español. No hemos encontrado trabajos que respondan a esta cuestión, a excepción de nuestro trabajo previo en el que observamos que parecen exis-

[1] El experimento consistía en la identificación, sin entrenamiento previo, de cuatro resúmenes de artículos de investigación lingüística en inglés.

tir, al menos, seis tipos de rasgos lingüísticos potencialmente discriminativos entre los textos humanos y los automáticos: el uso del léxico, la utilización de expresiones idiomáticas, el uso del orden sintáctico canónico SVO, de las estructuras comparativas y superlativas, el manejo de los marcadores discursivos y el uso de la puntuación (Alonso Simón, Gonzalo Gimeno, Fernández-Pampillón Cesteros, Fernández Trinidad y Escandell Vidal, 2023). El objetivo del presente estudio es verificar estadísticamente el carácter discriminativo del léxico, de los signos de puntuación y del orden sintáctico canónico entre textos humanos y textos generados por los modelos GPT-3.5 y GPT-4.

El uso del español como lengua de trabajo es relevante, fundamentalmente, porque es una de las lenguas más habladas (Fernández Vítores, 2023). Sin embargo, se dispone todavía de escasos recursos y aplicaciones de PLN. En este sentido, este estudio puede contribuir a aumentar el conocimiento y los recursos de PLN para el español.

Hemos organizado el artículo en siete secciones. En la sección 2 presentamos una síntesis de los trabajos previos, junto con sus principales conclusiones. En la sección 3 explicitamos las dos preguntas de investigación y las hipótesis de partida. En la sección 4 describimos el conjunto de datos utilizado para demostrar las hipótesis, el corpus ROBOT-TALK. En la sección 5 describimos la metodología. En la sección 6 discutimos los resultados obtenidos y la forma en que estos responden a las preguntas de investigación planteadas. Finalmente, en la sección 7 presentamos las conclusiones y futuras líneas de trabajo.

2. ESTADO DE LA CUESTIÓN

Hasta ahora, el problema de la distinción entre un texto generado por un LLM y otro escrito por un humano se ha formulado como una tarea de clasificación automática de autoría, habitualmente binaria *humano* vs. *automático*. Para crear estos clasificadores se han empleado aproximaciones estadísticas basadas en aprendizaje automático tradicional o en aprendizaje profundo (Jawahar, Abdul-Mageed y Lakshmanan, 2020). Los clasificadores basados en aprendizaje automático tradicional necesitan conocer cuáles son las características o rasgos discriminatorios que les permiten realizar dicha clasificación (Jurafsky y Martin, 2024). Estos rasgos deben ser definidos previamente por un especialista.

Ocurre que en los trabajos que hemos revisado, la selección de los rasgos discriminatorios se realiza de forma empírica, sin un estudio lingüístico sistemático previo de los textos (Berber Sardinha, 2024). Sencillamente, se realiza basándose en la intuición de los creadores de los clasificadores (que no son lingüistas), o en experimentos de identificación de textos humanos-automáticos por parte de personas que tampoco son especialistas en lingüística, o, en el mejor de los casos, en trabajos estilométricos[2] previos sobre análisis de autoría de textos humanos, como en Savoy (2020). Solo conocemos un caso en el que los humanos que participaron en la distinción entre textos humanos y automáticos eran lingüistas, aunque en dicho estudio no se buscaba la creación de un clasificador (Casal y Kessler, 2023).

Una vez definido el conjunto de rasgos, se construye el clasificador, se evalúa su eficacia mediante métricas como la precisión, la cobertura o el valor-F1 y, de forma iterativa, se va ajustando el conjunto de rasgos hasta obtener mejores valores de eficacia (Jawahar et al., 2020; Savoy, 2020). Entre los trabajos que siguen esta estrategia de construcción empírica del vector de rasgos se encuentran Pizarro (2019), Savoy (2020), Nguyen, Hatua y Sung (2023), Shijaku y Canhasi (2023), Zaitsu y Jin (2023), He, Mao y Liu (2024) y Berber Sardinha (2024).

Una segunda estrategia para la clasificación es la que se basa en clasificadores creados con grandes redes neuronales (aprendizaje profundo). En este caso, la red neuronal *aprende* por sí misma durante el entrenamiento cuáles son los rasgos característicos de los textos. El proble-

[2] La estilometría se basa en el análisis cuantitativo de los textos para tratar de identificar los patrones y características lingüísticas y paralingüísticas propios del estilo de cada autor.

ma es que los rasgos aprendidos están implícitamente incluidos en la configuración de la red neuronal (los pesos de las conexiones entre neuronas) y solo de forma indirecta es posible, en algunos casos, deducir estos rasgos. Esta aproximación es la que se ha seguido en los trabajos de Desaire, Chua, Isom, Jarosova y Hua (2023), Guo, Zhang, Wang, Jiang, Nie, Ding, Yue y Wu (2023), Mitrović, Andreoletti y Ayoub (2023) y Yu, Chen, Feng y Xia (2024).

En otros estudios, se opta por utilizar ambas estrategias (aprendizaje automático y aprendizaje profundo) para construir y evaluar varios clasificadores y encontrar, empíricamente, la mejor opción. Este es el caso de Uchendu, Le, Shu y Lee (2020), Fröhling y Zubiaga (2021), Desaire et al. (2023), Liao, Liu, Dai, Xu, Wu, Zhang, Huang, Zhu, Cai, Li, Liu y Li (2023), Ma, Liu, Yi, Cheng, Huang, Lu y Liu (2023) y Corizzo y Leal-Arenas (2023).

Respecto a la lengua de los textos estudiados, se observa que, de los dieciséis trabajos revisados, todos, excepto uno, se centran en el inglés. La excepción es el trabajo de Zaitsu y Jin (2023) que trabaja con el japonés, aunque en los trabajos de Pizarro (2019) y Corizzo y Leal-Arenas (2023), además del inglés, se incluye el español, y el chino en Guo et al. (2023).

Los LLM generativos utilizados en los trabajos revisados son, mayoritariamente, los modelos GPT en sus versiones 2, 3, 3.5 y 4. Uchendu et al. (2020) y Fröhling y Zubiaga (2021) eligieron GPT-2, GROVER (además de otros modelos); Zaitsu y Jin (2023) utilizaron GPT-3.5 y GPT-4, y el resto usó ChatGPT, el modelo GPT-3.5, o bien no especifican el modelo concreto, como en Corizzo y Leal-Arenas (2023), Desaire et al. (2023), Guo et al. (2023), Liao et al. (2023), Ma et al. (2023), Mitrović et al. (2023), Nguyen et al. (2023), Shijaku y Canhasi (2023), He et al. (2024), Yu et al. (2024) y Berber Sardinha (2024). Finalmente, Pizarro (2019) y Savoy (2020) eligieron los modelos generativos de los bots de Twitter.

La eficacia de los clasificadores construidos en estos trabajos presenta una métrica de evaluación macro-F1 de entre un 70% y un 98%. Sin embargo, es importante señalar que los valores altos se obtienen solamente cuando los corpus de evaluación pertenecen a la misma tipología textual que los textos utilizados en el entrenamiento. Cuando se aplica el clasificador a textos de tipología diferente de la utilizada para el entrenamiento, los valores se quedan en torno al 70%.

Finalmente, en la Tabla 1 se muestra una síntesis de los rasgos utilizados para la clasificación, junto con los autores que los han utilizado. Se ha querido separar los rasgos de puntuación del resto de los rasgos lingüísticos por su frecuente uso en los trabajos revisados y su relevancia para organizar bloques de información y sus relaciones en el texto escrito. Los rasgos de tipo lingüístico han sido utilizados por todos los autores; los rasgos de puntuación, por la mayoría de ellos y, otros tipos de características, paralingüísticas y psicolingüísticas, se han considerado solo de forma puntual.

Los rasgos de puntuación se utilizan en nueve de los dieciséis trabajos. La mayoría no especifica exactamente qué símbolos ha considerado; no obstante, es relevante que tres estudios coinciden en la importancia diferenciadora del estilo que ofrece el uso de la coma y del punto (Fröhling y Zubiaga, 2021; Corizzo y Leal-Arenas, 2023; Ma et al., 2023). En el estudio de Zaitsu y Jin (2023) se utiliza el posicionamiento de las comas como rasgo discriminatorio entre el estilo humano y el automático.

A partir de los valores de estos rasgos, algunos autores extraen conclusiones, no siempre coincidentes, sobre el diferente estilo de escritura en inglés de ChatGPT y el supuesto estilo de los humanos. Así, a nivel léxico, He et al. (2024), Liao et al. (2023), Guo et al. (2023) y Shijaku y Canhasi (2023) concluyen que la densidad de palabras de los humanos es mayor que la de ChatGPT. Por el contrario, Yu et al. (2024) y Berber Sardinha (2024) indican que ChatGPT puede proporcionar un vocabulario más diverso y hacer los textos más informativos. Corizzo y Leal-Arenas (2023) indican que las máquinas utilizan palabras muy frecuentes en los corpus mientras que, en sentido opuesto, Mitrović et al. (2023) sostienen que los textos de ChatGPT tienden a usar palabras poco comunes.

Tabla 1: ***Síntesis de rasgos distintivos de los textos humanos frente a automáticos***

Tipos de rasgos	**Autores**
1) Rasgos lingüísticos:	Todos
recuento de caracteres y palabras. Incluye: unigramas de caracteres o palabras, TTR (tokens únicos/ total-tokens), LD (palabras léxicas únicas/total-palabras), espacios en blanco, saltos de línea, longitud palabras, palabras únicas, palabras por oración, longitud media del texto, números, oraciones por texto, raíces únicas de palabras, frecuencia de palabras funcionales, palabras con mayúscula, errores gramaticales o de tipeo y número de párrafos.	Pizarro (2019), Savoy (2020), Uchendu et al. (2020), Fröhling y Zubiaga (2021), Corizzo y Leal-Arenas (2023), Desaire et al. (2023), Guo et al. (2023), Liao et al. (2023), Ma et al. (2023), Nguyen et al. (2023), Shijaku y Canhasi (2023), He et al. (2024) y Berber Sardinha (2024)
recuento de combinaciones de caracteres. Incluye: n-gramas de diferentes rangos.	Pizarro (2019)
recuento de combinaciones de palabras. Incluye: n-gramas de diferentes rangos.	Pizarro (2019), Fröhling y Zubiaga (2021), Corizzo y Leal-Arenas (2023), Nguyen et al. (2023), Shijaku y Canhasi (2023), He et al. (2024)
recuento de categorías léxicas. Incluye: entidades nombradas, adjetivos, adverbios, conjunciones, sustantivos, nombres propios, números, pronombres, verbos y preposiciones.	Fröhling y Zubiaga (2021), Guo et al. (2023), Liao et al. (2023), Ma et al. (2023), Nguyen et al. (2023), Shijaku y Canhasi (2023), Zaitsu y Jin (2023), He et al. (2024), Yu et al. (2024), Berber Sardinha (2024)
sintáctico-semánticos. Incluye: patrones argumentales y análisis de dependencias.	Guo et al. (2023), Liao et al. (2023), Yu et al. (2024), Berber Sardinha (2024)
semántico-pragmáticos. Incluye: opinión, enfáticos, atenuadores, intensificadores y coherencia semántica.	Corizzo y Leal-Arenas (2023), Guo et al. (2023), Liao et al. (2023), Ma et al. (2023), He et al. (2024), Berber Sardinha (2024)
pragmático-discursivos. Incluye: complejidad del párrafo, lógica de argumentación, partículas discursivas y legibilidad del texto.	Uchendu et al. (2020), Fröhling y Zubiaga (2021), Desaire et al. (2023), Liao et al. (2023), Ma et al. (2023), Mitrović et al. (2023), Nguyen et al. (2023), Berber Sardinha (2024)
2) Rasgos de puntuación. Incluye: paréntesis, guion corto, punto y coma, dos puntos, interrogación, comillas, coma y punto, entre otros.	Fröhling y Zubiaga (2021), Corizzo y Leal-Arenas (2023), Desaire et al. (2023), Guo et al. (2023), Ma et al. (2023), Nguyen et al. (2023), Shijaku y Canhasi (2023), Zaitsu y Jin (2023), Yu et al. (2024)
3) Rasgos paralingüísticos. Incluye: hipervínculos, menciones, *hashtags* y *emojis*.	Savoy (2020)
4) Rasgos psicolingüísticos. Incluye: procesos psicológicos y preocupaciones personales.	Uchendu et al. (2020)

A nivel sintáctico, se concluye que los textos humanos tienden a aplicar más modificadores (adjetivales, preposicionales y compuestos) para definir las palabras con precisión (Yu et al., 2024). También sugieren que las proporciones del objeto directo, el sujeto nominal y las palabras raíz[3] en los textos humanos son menores que las de los textos de ChatGPT (Yu et al., 2024). Finalmente, se ha observado que ChatGPT tiende a usar combinaciones sistemáticas de palabras, mientras que los humanos utilizan una mayor variedad de palabras individuales en contraposición a combinaciones de palabras (He et al., 2024).

A nivel pragmático-discursivo, Mitrović et al. (2023) concluyeron que ChatGPT tiende a expresarse de forma más impersonal y formal, y que carece de expresiones emocionales, frecuentemente utilizadas por los humanos. Yu et al. (2024) observaron que los humanos suelen prestar más atención a la estructura, la consistencia y la lógica del texto. He et al. (2024) concluyeron que los textos de ChatGPT parecen ser más sistemáticos y menos específicos en contenido que los textos humanos. Guo et al. (2023) determinaron que la frecuente coocurrencia de conjunciones, junto con sustantivos, verbos y preposiciones, indica que la estructura del texto y las relaciones de causa-efecto, progresión o contraste son claras. Finalmente, Berber Sardinha (2024) llevó a cabo un análisis lingüístico sistemático, considerando diversos dominios textuales, y observó que, en la conversación, ChatGPT utiliza aproximadamente la mitad de las características relacionadas con el grado de implicación del emisor en los textos que los humanos. Además, concluyó que ChatGPT no es capaz de captar la complejidad de la gramática del lenguaje hablado y que se basa en patrones del lenguaje escrito para generar los diálogos. También encontró que los textos generados automáticamente tienen 1,4 veces menos probabilidades de emplear referencias dependientes del contexto. Finalmente, observó que los textos generados por ChatGPT no suelen adoptar un tono persuasivo a diferencia de los textos humanos.

3. PREGUNTAS DE INVESTIGACIÓN E HIPÓTESIS

En consonancia con nuestro objetivo, las preguntas de investigación que nos planteamos en este trabajo son:

1) ¿podemos diferenciar mediante rasgos lingüísticos el estilo de escritura de GPT-3.5 de un posible estilo general humano en español?
2) ¿podemos diferenciar mediante rasgos lingüísticos el estilo de escritura de GPT-4 de un posible estilo general humano en español?

Entendemos por estilo general humano un posible conjunto de rasgos lingüísticos compartidos (si existen) por los autores humanos y que son claramente diferentes de los rasgos lingüísticos compartidos (si existen) por los sistemas de generación automáticos.

La hipótesis de partida es que los modelos GPT-3.5 (concretamente, GPT-3.5-Turbo) y GPT-4 tienen un estilo propio distinto de un estilo general humano y que, entre otros posibles, existen rasgos léxicos, ortográficos y sintácticos que permitirán diferenciarlos. La Tabla 2 recoge los dieciocho rasgos léxicos que hipotetizamos serán diferenciales entre los estilos de escritura humano y los dos robóticos. La Tabla 3 muestra los veintiún rasgos ortográficos potencialmente discriminativos.

[3] En análisis de dependencias, palabra principal de la que dependen directa o indirectamente las demás.

Tabla 2: ***Propuesta de rasgos léxicos diferenciadores de los estilos de escritura de GPT-3.5 y GPT-4 frente al humano***

Rasgos léxicos	**Valor**
Ratio Tipo de Token o *TTR* (*Type-Token Ratio*)	palabras únicas/total palabras
Densidad Léxica o *LD* (*Lexical Density*)	palabras léxicas únicas/total palabras
proporción de adjetivos	total adjetivos/total palabras
proporción de adjetivos únicos	total adjetivos únicos/total palabras
proporción de adverbios	total adverbios/total palabras
proporción de adverbios únicos	total adverbios únicos/total palabras
proporción de conjunciones	total conjunciones/total palabras
proporción de conjunciones únicas	total conjunciones únicas/total palabras
proporción de sustantivos	total sustantivos/total palabras
proporción de sustantivos únicos	total sustantivos únicos/total palabras
proporción de preposiciones	total preposiciones/total palabras
proporción de preposiciones únicas	total conjunciones únicas/total palabras
proporción de pronombres	total pronombres/total palabras
proporción de pronombres únicos	total pronombres únicos/total palabras
proporción de verbos	total verbos/total palabras
proporción de verbos únicos	total verbos únicos/total palabras
densidad de *hápax legómenon*	palabras con frecuencia 1/vocabulario único
densidad *dís legómenon*	palabras con frecuencia 2/vocabulario único

Tabla 3: ***Propuesta de rasgos ortográficos diferenciadores de los estilos de escritura de GPT-3.5 y GPT-4 frente al humano***

Rasgo	**Valor**
proporción de símbolos ortográficos total	total puntuación/total tokens
proporción de comillas (comillas rectas dobles o simples, comillas españolas, comillas inclinadas)	total comillas/total tokens
proporción de dos puntos	total dos puntos/total tokens
proporción de comas	total comas/total tokens
proporción de puntos suspensivos o la abreviatura ‘etc.’	total ps_etc/total tokens
proporción de corchetes de apertura	total carácter/total tokens
proporción de corchetes de cierre	total carácter/total tokens
proporción de guiones cortos	total carácter/total tokens
proporción de admiraciones de apertura	total carácter/total tokens
proporción de admiraciones de cierre	total carácter/total tokens

(Tabla 3, continúa en la página siguiente)

(Tabla 3, continúa de la página anterior)

Rasgo	**Valor**
proporción de paréntesis de apertura	total carácter/total tokens
proporción de paréntesis de cierre	total carácter/total tokens
proporción de puntos (punto y seguido o punto con cambio de línea)	total puntos/total tokens
proporción de barras inclinadas	total carácter/total tokens
proporción de interrogaciones de apertura	total carácter/total tokens
proporción de interrogaciones de cierre	total carácter/total tokens
proporción de llaves de apertura	total carácter/total tokens
proporción de llaves de cierre	total carácter/total tokens
proporción de símbolo de porcentaje '%'	total carácter/total tokens
proporción de punto y comas	total carácter/total tokens
proporción de otra puntuación (p. ej. guion normal, guion largo, marca de párrafo, signo + o asteriscos)	total otra puntuación/total tokens

Finalmente, los rasgos sintácticos discriminatorios pueden ser:

1) La proporción del número de oraciones por el total de palabras en el texto.
2) La proporción de oraciones con estructura canónica SVO.

4. EL CORPUS ROBOT-TALK

Para comprobar las hipótesis hemos utilizado una muestra del corpus ROBOT-TALK.[4] Se trata de un corpus comparable de textos en español (artículos científicos de lingüística, noticias y reseñas de cine) escritos por humanos y por cuatro LLM diferentes. La selección de los tres géneros se ha realizado de forma que estén representados textos conforme diferentes niveles de uso de lenguaje formulaico. Actualmente, consta de un total de 765 textos, la mitad son humanos y la otra mitad son textos comparables generados automáticamente por los LLM ChatGPT-3.5-turbo de OpenAI[5], ChatGPT-4 de OpenAI[6], Gemini de Google[7] y Mixtral-8x7B-Instruct-v0.1 de Mixtral AI[8]. El periodo de recogida de datos abarca desde julio de 2023 hasta abril de 2024. La Tabla 4 muestra el número de textos por autor y por dominio y la Tabla 5 recoge el número de tokens por autor y por dominio.

Desde el punto de vista de la clasificación automática, el corpus ROBOT-TALK es un corpus multiclase respecto a los cinco posibles autores: humano o uno de los cuatro modelos de lenguaje. En este sentido, para la tarea de detección de textos generados automáticamente, el corpus puede utilizarse, bien para la clasificación multiclase (atribución de autoría) con las etiquetas *humano*, *modelo1*, *modelo2*, *modelo3* y *modelo4*, bien para la clasificación binaria (Turing Test) con las etiquetas *humano* y *modelo*.

[4] El corpus ROBOT-TALK se ha creado en el marco del proyecto de investigación PID2022-140897OB-I00.

[5] https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/.

[6] https://openai.com/index/gpt-4/.

[7] https://gemini.google.com/app?hl=es-ES.

[8] https://mistral.ai/news/mixtral-of-experts/.

Tabla 4: ***Número de textos por autor y dominio del corpus ROBOT-TALK***

	Autor					**N.º de textos por dominio**
Dominio de los textos	**Humano**	**GPT-3.5-Turbo**	**GPT-4**	**Gemini**	**Mixtral**	
Artículos científicos	23	23	23	23	23	**115**
Noticias	65	65	65	65	65	**325**
Reseñas de cine	65	65	65	65	65	**325**
Número de textos	**153**	**153**	**153**	**153**	**153**	765

Tabla 5: ***Número de tokens por autor y dominio del Corpus ROBOT-TALK***

	Autor					**N.º de tokens por dominio**
Dominio de los textos	**Humano**	**GPT-3.5-Turbo**	**GPT-4**	**Gemini**	**Mixtral**	
Artículos científicos	33.189	21.833	23.508	27.791	33.043	**139.364**
Noticias	30.003	24.821	30.811	25.084	29.013	**139.732**
Reseñas de cine	24.667	21.969	24.300	23.928	24.541	**119.405**
Número de tokens	**87.859**	**68.623**	**78.619**	**76.803**	**86.597**	398.501

Hasta donde sabemos, por el momento, no existe otro corpus comparable de textos escritos por hablantes nativos de español, cuya autoría humana haya sido rigurosamente comprobada[9], y por textos equivalentes y no híbridos (combinaciones de oraciones escritas por humanos y oraciones producidas por modelos generativos), generados por GPT-3.5-Turbo, GPT-4, Gemini y Mixtral-8x7B-Instruct-v0.1 que, además, contemplen los tres dominios (artículos, noticias y reseñas).

5. MÉTODO

Para comprobar si los rasgos lingüísticos definidos en la hipótesis asumen valores que resulten ser característicos de los dos modelos generativos considerados en este estudio, GPT-3.5 y GPT-4, y de un estilo humano común de escritura, hemos utilizado los análisis estadísticos de varianza y covarianza que nos permiten conocer si las medias obtenidas para cada rasgo lingüístico son significativamente diferentes entre los autores.

5.1. *Muestra*

La muestra utilizada tiene 180 textos del corpus ROBOT-TALK, elegidos al azar y de forma equilibrada respecto a la tipología (20 artículos, 20 noticias y 20 reseñas) y a cada tipo de autor: humano, GPT-3.5 y GPT-4 (Tabla 6).

[9] Sarvazyan, González, Franco-Salvador, Rangel, Chulvi y Rosso (2023) utilizan textos en español en el corpus AuTexTification, pero (1) no se ha comprobado la autoría humana de los textos etiquetados como humanos y (2) no es un corpus comparable.

Tabla 6: ***Número de textos por autor y dominio de la muestra utilizada***

	Autor			**N.º de textos por dominio**
Dominio de los textos	**Humano**	**GPT-3.5-Turbo**	**GPT-4**	
Artículos científicos	20	20	20	**60**
Noticias	20	20	20	**60**
Reseñas de cine	20	20	20	**60**
Número de textos	**60**	**60**	**60**	180

5.2. *Herramientas*

Hemos utilizado la herramienta de análisis textual Sketch Engine (Kilgarriff, Baisa, Bušta, Jakubíček, Kovář, Michelfeit, Rychlý y Suchomel, 2014) para construir el corpus de muestra y obtener automáticamente los valores de cada una de las variables.[10] Los datos se han almacenado en hojas de cálculo. El análisis estadístico se ha realizado con el programa SPSS.

5.3. *Hipótesis*

H1.1: existen diferencias significativas entre los valores medios de las variables evaluadas en los textos de GPT-3.5 y humanos.
H1.2: existen diferencias significativas entre los valores medios de las variables evaluadas, estudiadas por dominios (artículos, noticias y reseñas), en los textos de GPT-3.5 y humanos.
H2.1: existen diferencias significativas entre los valores medios de las variables evaluadas en los textos de GPT-4 y humanos.
H2.2: existen diferencias significativas entre los valores medios de las variables evaluadas, estudiadas por dominios (artículos, noticias y reseñas), en los textos de GPT-4 y humanos.

5.4. *Procedimiento*

Hemos realizado cuatro análisis comparando los textos dos a dos:

1) GPT-3.5 vs. humano en todos los textos
2) GPT-3.5 vs. humano en cada tipo de texto
3) GPT-4 vs. humano en todos los textos
4) GPT-4 vs. humano en cada tipo de texto

Los análisis realizados han sido:

1) en las comparaciones con todos los textos, prueba de t-Student pareada, si la muestra presenta una distribución normal tras un test Shapiro-Wilk y rangos signados de Wilcoxon, si la distribución se aleja de la normalidad.
2) en las comparaciones múltiples respecto a cada tipo de texto utilizamos ANOVA y test de Duncan, en caso de normalidad, y ANOVA y test de Kruskal-Wallis, si la muestra se aleja de la normalidad.

[10] http://www.sketchengine.eu.

6. RESULTADOS Y DISCUSIÓN

Los resultados detallados de las diferencias obtenidas para los 41 rasgos estudiados entre GPT-3.5 y humanos y entre GPT-4 y humanos pueden encontrarse en la plataforma Open Science Framework.[11] Resumimos y discutimos los resultados más sobresalientes en las tres subsecciones siguientes.

6.1. *Diferencias en los rasgos léxicos*

Sin tener en cuenta las distinciones de dominio al que pertenecen los textos (Figura 1):

1) Existe una diferencia moderadamente significativa en el TTR en favor de los humanos, que indica que los textos humanos tienen mayor riqueza léxica que los textos de GPT-3.5 y GPT-4 (+2,3% y +2,8%).
2) En relación con la LD podemos observar que no hay diferencias significativas, por lo que el grado de informatividad de los textos es similar entre los textos LLM y los humanos. Curiosamente, GPT-4 utiliza, con una diferencia significativa baja, menos palabras léxicas únicas que los humanos, mientras que esta diferencia no es significativa entre GPT-3.5 y los humanos.
3) Las categorías léxicas de adjetivo, adverbio y pronombre, tanto en su frecuencia de uso como en el vocabulario (número de palabras únicas), presentan diferencias muy significativas entre los LLM (GPT-3.5 y GPT-4) y los humanos (Figura 1). Así, los textos humanos presentan un mayor uso y variedad de adverbios y pronombres mientras que, curiosamente, los textos generados por GPT-3.5 y GPT-4 exhiben más adjetivos y adjetivos únicos en comparación con los textos humanos. Además, GPT-3.5 se diferencia de los humanos en el uso de los adverbios únicos y las conjunciones únicas de forma muy significativa. Sin embargo, esto no ocurre en los textos de GPT-4.
4) Hay seis rasgos en los que las diferencias entre LLM y humanos no son significativas o son poco significativas: nombres y nombres únicos, preposiciones y preposiciones únicas y verbos y verbos únicos.
5) Las diferencias entre GPT-4 y los humanos son ligeramente menos marcadas en algunos rasgos que entre GPT-3.5 y los humanos. En español, y respecto a estos rasgos, parece que el estilo de GPT-4 es algo más cercano al humano que el de GPT-3.5. Esta observación se mantiene también en los tres dominios (Figuras 2, 3 y 4).

Las diferencias observadas entre los rasgos léxicos en las comparaciones entre GPT-3.5 vs. humanos y GPT-4 vs. humanos también se confirman en cada uno de los dominios (Figuras 2, 3 y 4). Esto podría indicar que, además de diferenciarse de los humanos, GPT-3.5 y GPT-4 podrían tener sus propios estilos de escritura.

Analizando los resultados por dominios observamos que el dominio de las reseñas es el más productivo, pues presenta el mayor número de rasgos diferenciales, doce, que permitirían distinguir el estilo de los textos de los LLM de los humanos (Figura 4). Resulta llamativo que, en el dominio más estructurado y formulaico de los artículos lingüísticos, exista también un número importante de rasgos distintivos, once para GPT-3.5 y ocho para GPT-4 (Figura 2). Sin embargo, el dominio de las noticias (Figura 3), con un nivel intermedio de restricciones estilísticas, ha resultado ser el menos adecuado para caracterizar el estilo humano o robótico (GPT-3.5 y GPT-4).

[11] https://osf.io/3ahb7/?view_only=076c6b452d68430e812826d91accd9d0

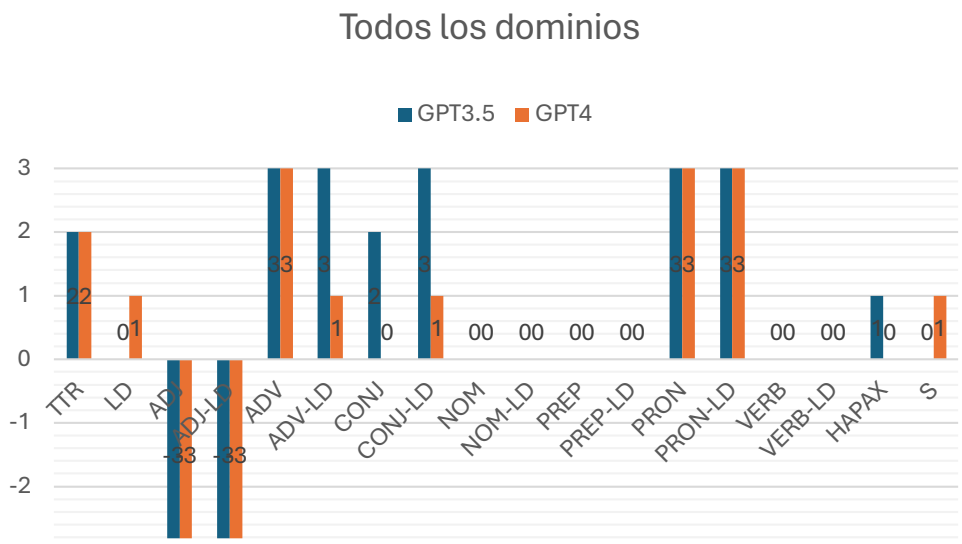

Figura 1: ***Nivel de significancia de la diferencia entre los LLM GPT-3.5 y GPT-4 y los humanos para los rasgos léxicos analizados en todos los dominios***

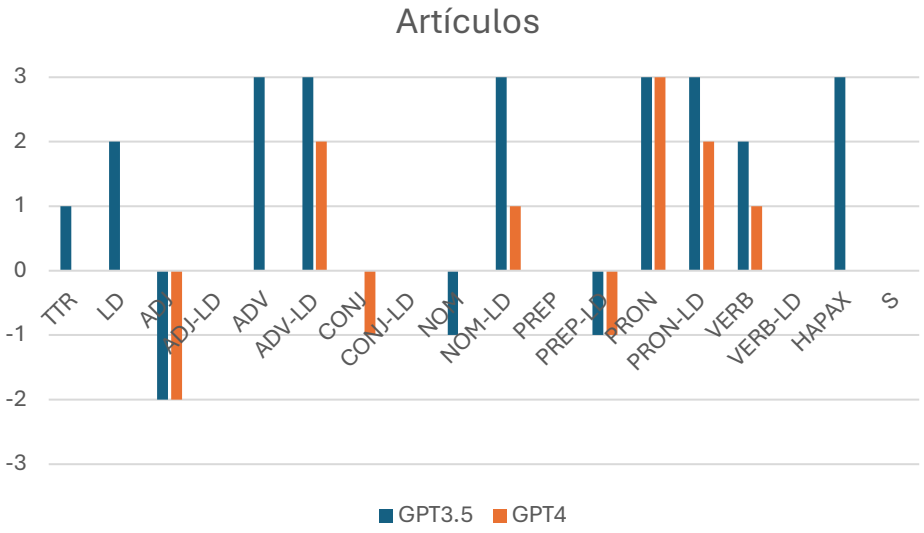

Figura 2: ***Nivel de significancia de la diferencia entre los LLM GPT-3.5 y GPT-4 y los humanos para los rasgos léxicos analizados en los textos pertenecientes al dominio de los artículos científicos***

Centrándonos en el dominio de los artículos científicos, cabe destacar tres resultados (Figura 2):

1) En este dominio, GPT-4 presenta una riqueza de vocabulario semejante a la de los humanos, por los valores nulos que presenta en la diferencia modelo vs. humano en los rasgos TTR y LD, y superior a la de GPT-3.5.
2) En los rasgos adjetivos, conjunciones únicas (sólo GPT4), nombres (sólo GPT3.5) y preposiciones únicas los modelos presentan una mayor riqueza de uso o de vocabulario que los humanos.
3) Los humanos utilizan significativamente más hápax (palabras únicas) que GPT-3.5; sin embargo, no se han hallado diferencias para este rasgo en la comparación GPT-4 y humanos.

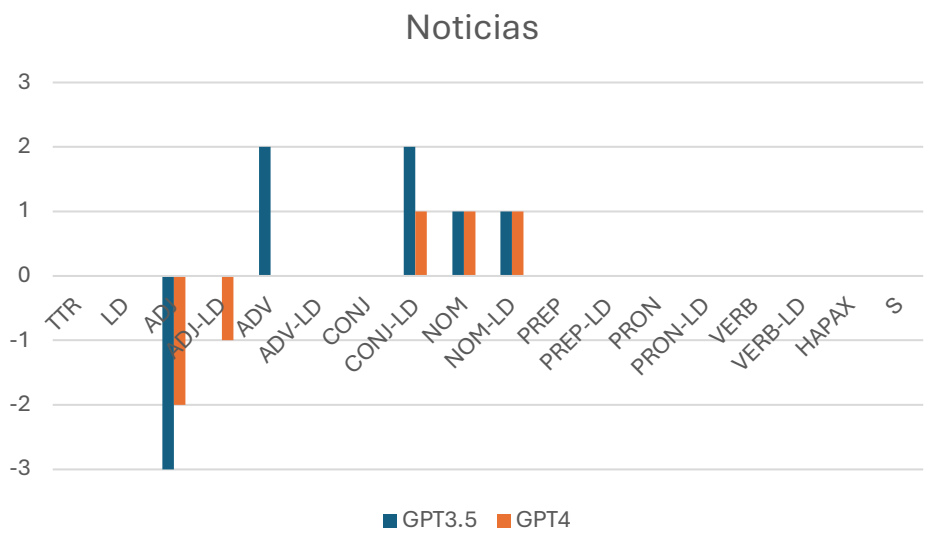

Figura 3: ***Nivel de significancia de la diferencia entre los LLM GPT-3.5 y GPT-4 y los humanos para los rasgos léxicos analizados en los textos pertenecientes al dominio de las noticias***

En el dominio de noticias, es reseñable que GPT4 es el modelo en el que las diferencias con los humanos tienen menos significancia respecto a GPT3.5 (Figura 3). Podríamos decir que, en este dominio, es en el que existe más parecido entre GPT4 y los humanos. Los resultados son coherentes con los obtenidos en el ámbito de los artículos científicos.

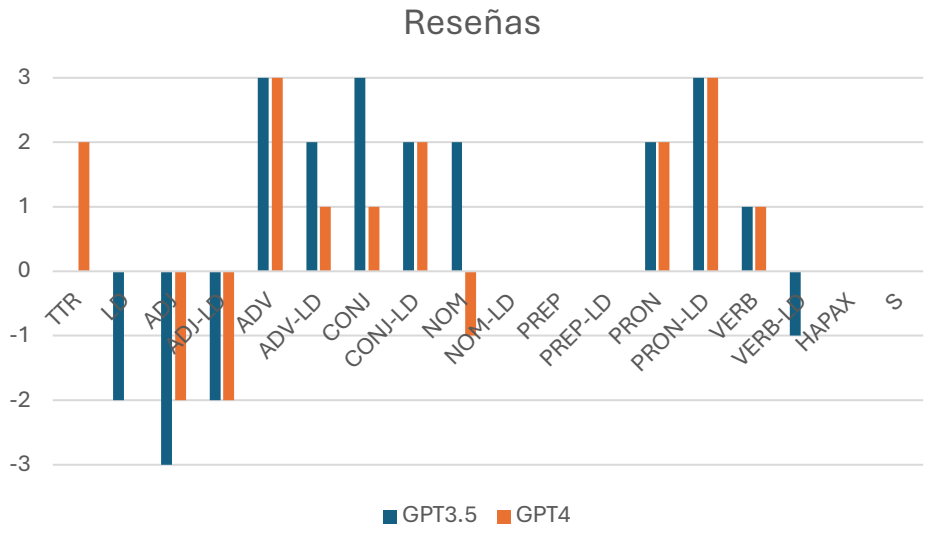

Figura 4: ***Nivel de significancia de la diferencia entre los LLM GPT-3.5 y GPT-4 y los humanos para los rasgos léxicos analizados en los textos pertenecientes al dominio de las reseñas de cine***

En el dominio de las reseñas, ámbito en el que encontramos más rasgos diferenciales, constatamos que:

1) Al contrario de lo que observábamos para los artículos, GPT-3.5 presenta una densidad léxica mayor en comparación con los humanos.

2) Los modelos presentan una mayor riqueza en el uso de adjetivos y adjetivos únicos (también ocurre en artículos y noticias), nombres (solo GPT-4) y verbos únicos (solo GPT-3.5) respecto a los textos humanos.
3) Existen diferencias entre los rasgos distintivos entre GPT-3.5 vs. humanos y GPT-4 vs. humanos. Esto podría indicar que GPT-3.5 y GPT-4 podrían tener sus propios estilos de escritura diferentes entre sí, además de diferenciarse, en conjunto, del estilo de escritura humano.

6.2. *Diferencias en los rasgos de puntuación*

Considerando el conjunto de todos los textos, observamos que, en español y en el corpus ROBOT-TALK, los humanos utilizan, en total, más signos de puntuación que GPT-3.5 y GPT-4 —de forma muy significativa en GPT-3.5 y de forma moderada en GPT-4— excepto en la utilización del punto. En este caso, los modelos usan significativamente más puntos que los humanos (Figura 5 y Tabla 7).

Específicamente, encontramos diferencias muy significativas en diez signos de puntuación: comillas, dos puntos, comas, guion corto, paréntesis de apertura y de cierre, punto, barra inclinada, punto y coma y otros signos que incluyen guion normal, guion largo, marca de párrafo, signo + y asteriscos. Esta importante cantidad de rasgos de puntuación diferencial es un resultado relevante de cara a poder identificar los estilos de los modelos frente a los humanos.

También encontramos siete rasgos que parecen no ser relevantes, entre los que se encuentran, de forma inesperada, las admiraciones. Sin embargo, conviene tener en cuenta que, respecto a cuatro de ellos, apertura y cierre de los corchetes y las llaves, no disponemos de datos porque no aparecen en los textos del corpus. Respecto a los otros tres (apertura y cierre de admiraciones y tanto por ciento), no disponemos de datos suficientes para poder extraer conclusiones.

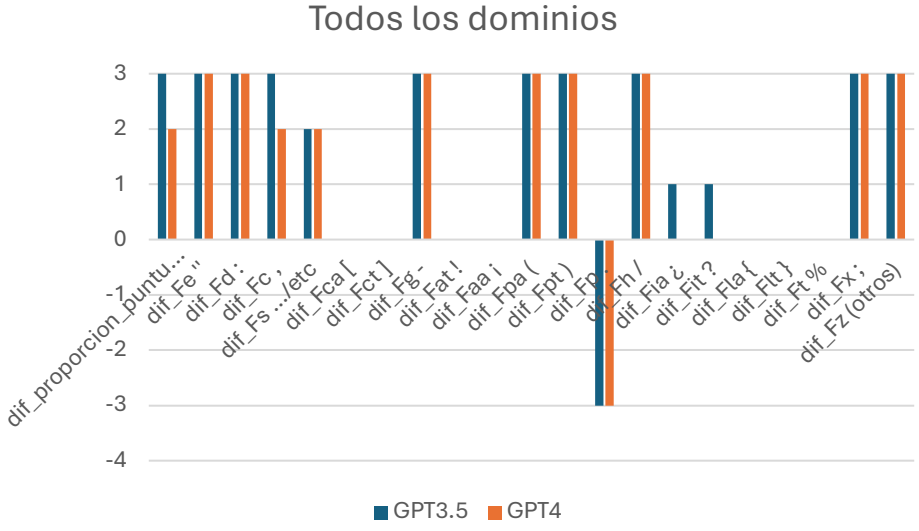

Figura 5: ***Nivel de significancia de la diferencia entre los LLM GPT-3.5 y GPT-4 y los humanos para los rasgos de puntuación analizados en todos los dominios***

Tabla 7: ***Nivel de significancia de la diferencia entre GPT-3.5 y GPT-4 respecto a los humanos para los rasgos de puntuación analizados en todos los dominios***

	GPT3.5	GPT4
dif_proporcion_puntuacion	3	2
dif_Fe ‘’	3	3
dif_Fd :	3	3
dif_Fc ,	3	2
dif_Fs …/etc	2	2
dif_Fca [	0	0
dif_Fct]	0	0
dif_Fg -	3	3
dif_Fat !	0	0
dif_Faa ¡	0	0
dif_Fpa (	3	3
dif_Fpt)	3	3
dif_Fp .	-3	-3
dif_Fh /	3	3
dif_Fia ¿	1	0
dif_Fit ?	1	0
dif_Fla {	0	0
dif_Flt }	0	0
dif_Ft %	0	0
dif_Fx ;	3	3
dif_Fz (otros)	3	3

Considerando cada dominio de forma independiente, en los artículos (Figura 6) se corroboran prácticamente todos los resultados hallados a nivel global. Las pequeñas discrepancias deberían estudiarse de forma cualitativa.

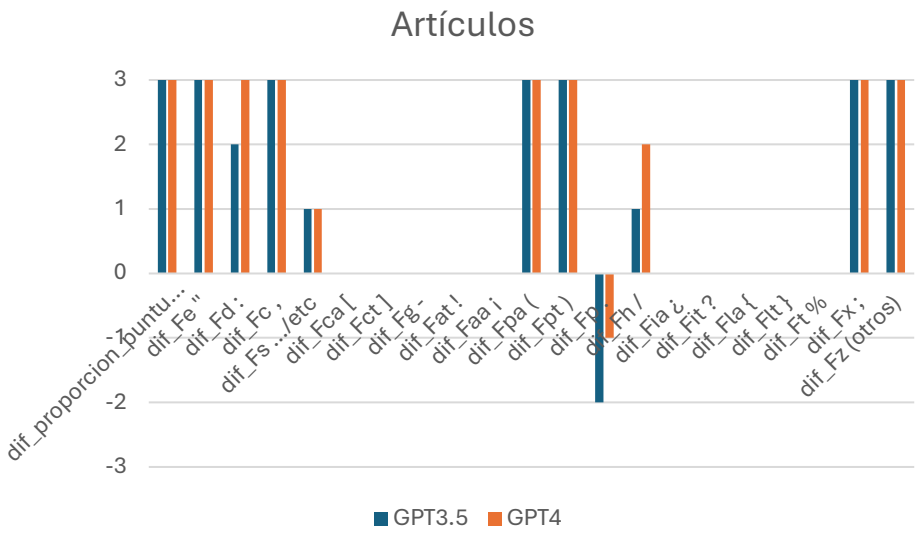

Figura 6: ***Nivel de significancia de la diferencia entre los LLM GPT-3.5 y GPT-4 y los humanos para los rasgos de puntuación analizados en todos los textos pertenecientes al dominio de los artículos científicos***

En las noticias ocurre, igual que sucede con los rasgos léxicos, que los rasgos diferenciales de puntuación son sensiblemente menores que en los otros dos dominios, corroborando que los estilos de GPT-3.5 y, especialmente el de GPT-4, son más similares al estilo humano. Únicamente el uso de comillas, de paréntesis y de puntos pueden ayudar a distinguirlos para GPT-4 y, además, las comas para GPT-3.5 (Figura 7).

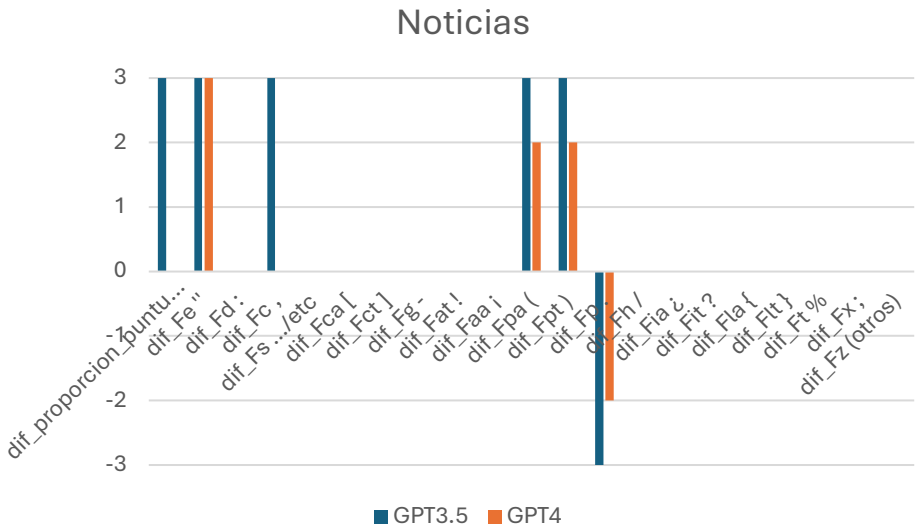

Figura 7: ***Nivel de significancia de la diferencia entre los LLM GPT-3.5 y GPT-4 y los humanos para los rasgos de puntuación analizados en los textos pertenecientes al dominio de las noticias***

Finalmente, y a diferencia de los rasgos léxicos, las reseñas es el dominio donde los signos de puntuación son menos relevantes para identificar los estilos de GPT-3.5 y GPT-4 frente a los humanos. Se siguen manteniendo como rasgos relevantes el uso de las comillas, los paréntesis y los puntos (Figura 8).

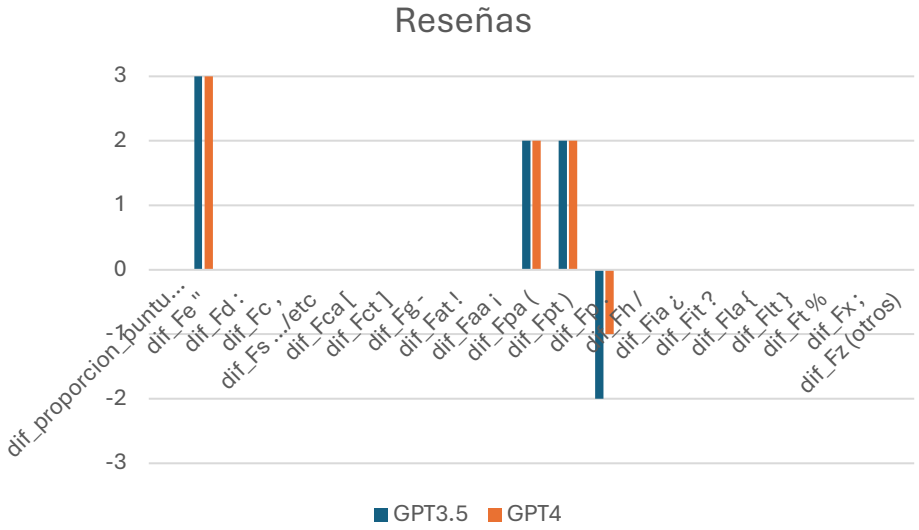

Figura 8: ***Nivel de significancia de la diferencia entre los LLM GPT-3.5 y GPT-4 y los humanos para los rasgos de puntuación analizados en los textos pertenecientes al dominio de las reseñas de cine***

En definitiva, en el análisis por dominios, es en los artículos donde se hallan las mayores diferencias entre los LLM y los humanos en cuanto al uso de la puntuación. Además, parece que los signos más informativos son las comillas, los paréntesis y los puntos: los humanos utilizan más comillas y paréntesis que GPT-3.5 y que GPT-4 y estos modelos utilizan más puntos que los humanos.

6.3. *Diferencias en los rasgos sintácticos*

Considerando todos los dominios, se observa que la proporción del número de oraciones por texto es significativamente mayor en los textos generados por GPT-3.5 y GPT-4 que en los textos humanos (Figura 9). El uso del orden canónico SVO es mayor en GPT-3.5 que en los humanos, aunque con una significancia baja y no se observan diferencias entre GPT-4 y los humanos. El análisis por dominios confirma estos resultados, siendo más relevante en los artículos que en las noticias y reseñas. De forma más detallada, el número de oraciones por texto es significativamente mayor en GPT-3.5 que en los humanos en todos los dominios mientras que en GPT4 la significancia es baja en las noticias y reseñas, La diferencia en la proporción del uso del orden canónico SVO tiene una significancia baja en los artículos y reseñas y no se observan diferencias en las noticias. Así pues, el único rasgo sintáctico relevante parece ser la proporción de oraciones en los textos (Figura 10).

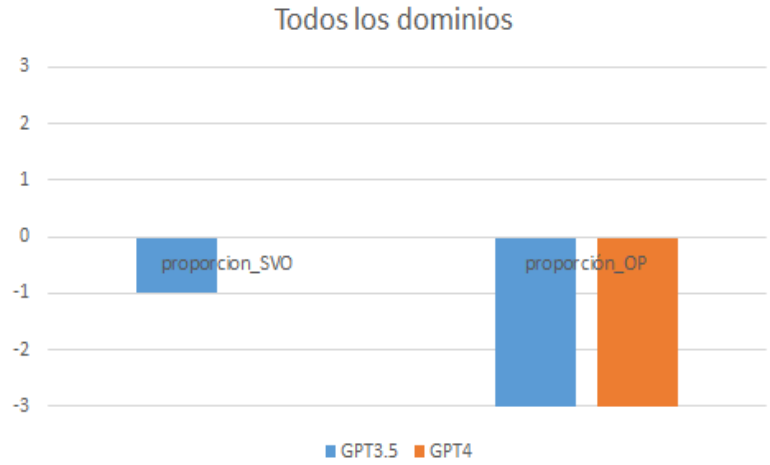

Figura 9: ***Nivel de significancia de la diferencia entre los LLM GPT-3.5 y GPT-4 y los humanos para los rasgos sintácticos analizados en todos los dominios***

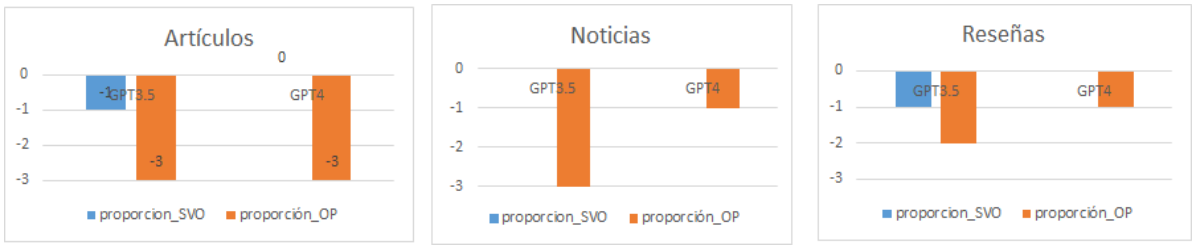

Figura 10: ***Nivel de significancia de la diferencia entre los LLM GPT-3.5 y GPT-4 y los humanos para los rasgos sintácticos analizados por dominios***

7. CONCLUSIONES, TRABAJO ACTUAL Y FUTURO

La pregunta de investigación que motiva este trabajo es si es posible verificar que, en español, los modelos GPT-3.5 y GPT-4 presentan un estilo de escritura característico identificable mediante rasgos lingüísticos y, si este estilo es diferente de un posible estilo general humano. Los resultados muestran que en 17 de los 41 rasgos analizados se exhiben diferencias significativas muy altas al comparar textos generados por GPT-3.5 y por GPT-4 con textos humanos.

Concretamente, se trata de los rasgos léxicos de uso de adjetivos y adjetivos únicos que es mayor en los LLM que en los humanos, así como de adverbios y adverbios únicos y de pronombres y pronombres únicos que es mayor en los humanos que en los LLM (aunque en GPT-4 el rasgo de adverbios únicos tiene una significancia más baja). Los rasgos de puntuación, uso de puntuación total, paréntesis, punto, coma, dos puntos, comillas, guion corto, barra inclinada, punto y coma y otros signos, y el rasgo sintáctico del número de oraciones por texto también han resultado ser caracterizadores del estilo de escritura de los LLM estudiados. En este caso, estos rasgos se presentan en mayor cantidad en los textos humanos que en los generados por GPT-3.5 y GPT4 excepto en el caso del punto que es a la inversa: los LLM utilizan más puntos que los humanos, lo que es congruente con el uso de más oraciones por texto.

Aunque el tamaño de la muestra escogida n=180 (60 textos por cada *autor* repartidos en 20 textos por dominio) proporciona suficiente confianza en los resultados obtenidos para todos los dominios, hemos encontrado algunos resultados referidos a cada tipo textual (artículos científicos, noticias y reseñas) que no son siempre concluyentes, pero que, por su interés, merecerían comprobarse repitiendo el estudio con una muestra aumentada. Además, es necesario confirmar con una muestra más amplia aquellas diferencias que resultaron ser significativas de forma moderada o baja. Sería interesante también añadir otras variables lingüísticas para continuar verificando las diferencias de estilo entre los LLM y los humanos. En este sentido, este estudio cuantitativo se está completando actualmente con un análisis cualitativo del uso de estos rasgos diferenciadores y con la búsqueda de otros posibles rasgos distintivos.

Los rasgos explorados en este estudio habían sido utilizados en trabajos anteriores para la construcción de clasificadores automáticos que diferencian texto humano de texto automático, fundamentalmente en lengua inglesa. Sin embargo, hasta el momento no se había verificado de forma intrínseca el grado de informatividad de éstos. Tampoco se había llevado a cabo un estudio de estas características en lengua española. Verificar este resultado es relevante para mejorar los métodos de detección de textos generados por modelos, en los que se apliquen soluciones informáticas y también lingüísticas, especialmente de lingüística forense, en los que interesa conocer quién es el autor de textos generados con fines maliciosos. Asimismo, creemos que contribuirá a mejorar la construcción de clasificadores automáticos y permitiría entrenar de forma más eficaz estos sistemas, mediante la eliminación de los rasgos que no pertenecen a los estilos de escritura de los modelos. Esta conclusión, en todo caso, todavía debe corroborarse empíricamente, por ejemplo, con la construcción de los clasificadores.

Otro resultado relevante de este estudio es que parece que GPT-3.5 y GPT-4 podrían tener sus propios estilos de escritura independientes. Esto nos lleva a pensar que podríamos hablar de que los modelos de lenguaje tienen un idiolecto propio, de manera semejante a lo que ocurre con autores humanos. Definir y verificar la existencia de idiolectos de los LLM requiere de la aplicación de modelos y métodos cualitativos y cuantitativos de naturaleza lingüística.

Finalmente, pensamos que las aportaciones de este trabajo constituyen un paso importante en el esfuerzo conjunto para entender mejor la generación automática de textos.

AGRADECIMIENTOS

Esta publicación es parte del proyecto de I+D+i Proyecto ROBOT-TALK PID2022-140897OB-I00 financiado por MCIN/AEI/10.13039/501100011033/ y FEDER/UE.

Agradecemos a Ricardo García Mata del Departamento de Apoyo Investigación (Servicios Informáticos UCM) su asesoramiento técnico para el análisis estadístico de los datos y al Dr. Miguel Jiménez-Bravo sus orientaciones en los experimentos preliminares. Finalmente, agradecemos las sugerencias y comentarios de los revisores anónimos.

REFERENCIAS

Alonso Simón, L., Gonzalo Gimeno, J. A., Fernández-Pampillón Cesteros, A. M.ª, Fernández Trinidad, M. y Escandell Vidal, M.ª V. (2023). Using Linguistic Knowledge for Automated Text Identification. En M. Montes y Gómez et al. (Eds.), *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023)*. Jaén, España, 26 de septiembre. https://ceur-ws.org/Vol-3496/autextification-paper17.pdf

Berber Sardinha, T. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics, 4*(1). https://doi.org/10.1016/j.acorp.2023.100083

Cañete, J., Chaperon, G., Fuentes, R., Ho, J-H., Kang, H. y Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. *arXiv:2308.02976v1*. https://doi.org/10.48550/arXiv.2308.02976

Cardenuto, J. P., Yang, J., Padilha, R., Wan, R., Moreira, D., Li, H., Wang, S., Andaló, F., Marcel, S. y Rocha, A. (2023). The Age of Synthetic Realities: Challenges and Opportunities. *APSIPA Transactions on Signal and Information Processing, 12*(1), 1–62. https://doi.org/10.1561/116.00000138

Casal, J. E. y Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics, 2*(3). https://doi.org/10.1016/j.rmal.2023.100068

Corizzo, R. y Leal-Arenas, S. (2023). A Deep Fusion Model for Human vs. Machine-Generated Essay Classification. En D. Wang y T. Toyoizumi (Eds.), *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. Gold Coast, Australia, 18-23 de junio. https://doi.org/10.1109/IJCNN54540.2023.10191322

Crothers, E. N., Japkowicz, N. y Viktor, H. L. (2023). Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *arXiv:2210.07321*, Oct. 2023. https://doi.org/10.1109/ACCESS.2023.3294090

Desaire, H., Chua, A. E., Isom, M., Jarosova, R. y Hua, D. (2023). Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science, 4*(6). https://doi.org/10.1016/j.xcrp.2023.101426

Fernández Vítores, D. (2023). El español: una lengua viva. Informe 2023. En C. Pastor Villalba (dir.), Instituto Cervantes (coord.), *El español en el mundo. Anuario del Instituto Cervantes 2023* (pp. 19-142). Madrid: Instituto Cervantes.

Fröhling, L. y Zubiaga, A. (2021). Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science, 7*, 1–23. https://doi.org/10.7717/PEERJ-CS.443

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J. y Wu, Y. (2023). How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation y Detection. *arXiv:2301.07597v1*. https://doi.org/10.48550/arXiv.2301.07597

Hadi, M. U., Al-Tashi, O., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M., Akhtar, N., Wu, J. y Mirjalili, S. (2023). *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects.* TechRxiv. https://doi.org/10.36227/techrxiv.23589741.v4

He, Z., Mao, R. y Liu, Y. (2024). Predictive model on detecting ChatGPT responses against human responses. *Applied and Computational Engineering, 44*(1), 18–25. https://doi.org/10.54254/2755-2721/44/20230078

Jawahar, G., Abdul-Mageed, M. y Lakshmanan, L. V. S. (2020). Automatic Detection of Machine Generated Text: A Critical Survey. En D. Scott, N. Bel, y C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 2296–2309). Barcelona: International Committee on Computational Linguistics. *arXiv:2011.01314.* https://doi.org/10.48550/arXiv.2011.01314

Jurafsky, D. y Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed. draft)*. Stanford University. Recuperado de https://web.stanford.edu/~jurafsky/slp3/

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. y Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Huang, X., Zhu, D., Cai, H., Li, Q., Liu, T. y Li, X. (2023). Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantitative Study. *JMIR Medical Education, 9*(1). https://doi.org/10.2196/48904

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W. y Liu, X. (2023). AI vs. Human-Differentiation Analysis of Scientific Content Generation. *arXiv:2301.10416v2.* https://doi.org/10.48550/arXiv.2301.10416

Maloyan, N., Nutfullin, B. y Ilyushin, E. (2022). DIALOG-22 RuATD Generated Text Detection. En *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"* (2022) (pp. 396–401). Moscú: RSUH. *arXiv.2206.08029.* https://doi.org/10.48550/arXiv.2206.08029

Mitrović, S., Andreoletti, D. y Ayoub, O. (2023). ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text. *arXiv:2301.13852v1.* https://doi.org/10.48550/arXiv.2301.13852

Nguyen, T. T., Hatua, A. y Sung, A. H. (2023). How to Detect AI-Generated Texts? En *IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 0464–0471). *IEEE;* Nueva York. https://doi.org/10.1109/UEMCON59035.2023.10316132

Pavlyshenko, B. M. (2022). Methods of Informational Trends Analytics and Fake News Detection on Twitter. *arXiv:2204.04891v1*. https://doi.org/10.48550/arXiv.2204.04891

Pizarro, J. (2019). Using n-grams to detect bots on Twitter, notebook for PAN at CLEF 2019. En L. Cappellato, N. Ferro, D. E. Losada, and H. Müller (Eds.), *CLEF 2019 Labs and Workshops, Notebook Papers*. https://ceur-ws.org/Vol-2380/paper_183.pdf

Radford, A., Narasimhan, K., Salimans, T. y Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *OpenAI Blog*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Sarvazyan, A. M., González, J. Á., Franco-Salvador, M., Rangel, F., Chulvi, B. y Rosso, P. (2023). Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains. *Procesamiento del Lenguaje Natural, 71*, 275–288. https://doi.org/10.26342/2023-71-21

Savoy, J. (2020). *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*. Neuchatel: Springer. https://doi.org/10.1007/978-3-030-53360-1

Shijaku, R. y Canhasi, E. (2023). *ChatGPT Generated Text Detection*. Artículo presentado en The Fifteenth International Conference on Future Computational Technologies and Application. Future Computing 2023. Nice, Saint-Laurent-du-Var, France. 26-30 de junio de 2023. https://doi.org/10.13140/RG.2.2.21317.52960

Uchendu, A., Le, T., Shu, K. y Lee, D. (2020). Authorship attribution for neural text generation. En B. Webber et al. (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 8384–8395. Online. https://doi.org/10.18653/v1/2020.emnlp-main.673

Uchendu, A., Le, T. y Lee, D. (2023). Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter, 25*(1), 1–18. https://doi.org/10.1145/3606274.3606276

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. y Polosukhin, I. (2017). Attention Is All You Need. En U. von Luxburg et al. (Eds.), *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems,* (pp. 6000–6010). Long Beach, California, Estados Unidos. https://doi.org/10.5555/3295222.3295349

Yu, P., Chen, J., Feng, X. y Xia, Z. (2024). CHEAT: A Large-scale Dataset for Detecting ChatGPT-writtEn AbsTracts. *arXiv:2304.12008v2*. https://doi.org/10.48550/arXiv.2304.12008

Zaitsu, W. y Jin, M. (2023). Distinguishing ChatGPT(-3.5, -4)-generated and human-written papers through Japanese stylometric analysis. *PLoS ONE, 18*(8). https://doi.org/10.1371/journal.pone.0288453