Propuesta recibida en 15 de octubre 2017 y aceptada para publicación en 3 de diciembre 2017.

Las leyes estadísticas del lenguaje como marcador de estilo para la detección de textos robóticos en español

Statistical Laws of Language as a Stylometric Fingerprint for Machine-Generated Text Detection in Spanish

Isabel Moyano-Moreno
Instituto Universitario de Investigación en Lingüística Aplicada, Universidad de Cádiz

Resumen

Este trabajo evalúa la capacidad discriminativa de un conjunto de descriptores lingüísticos y leyes estadísticas (Zipf, Heaps y el principio de brevedad) para diferenciar entre autoría humana y robótica en artículos científicos en español. A partir del corpus ROBOT-TALK (Alonso Simón et al., 2024), se construye un conjunto pareado de 90 artículos, cada uno con su versión humana y sus seis equivalentes robóticos (Bard, Claude 3.5 Sonnet, DeepSeek-V3, GPT-3.5-turbo, GPT-4, Mixtral-8x7B-Instruct). Se extraen 15 variables estilométricas que no dependen del vocabulario concreto del texto, evaluadas mediante validación cruzada agrupada. Las variables de longitud oracional —asociadas a la dimensión sintáctica del principio de brevedad— son el grupo de rasgos individual más discriminativo; el conjunto completo alcanza F1=0,943, comparable al Delta de Burrows (2002) pero sin recurrir a ninguna forma léxica concreta. La tasa de falsos positivos sobre textos humanos es inferior al 4 % (FPR=0,033; SVM). Los modelos más recientes (Claude 3.5 Sonnet, DeepSeek-V3, GPT-4) resultan los más próximos al perfil estadístico humano, mientras que diferenciar los LLMs entre sí es más exigente (F1 macro=0,699). Estos resultados sugieren que los descriptores lingüísticos de naturaleza oracional operan como marcadores estilométricos interpretables e independientes del vocabulario para el análisis de autoría en la era de los LLMs.

Palabras clave

análisis de autoría, estilometría, leyes estadísticas del lenguaje, lingüística cuantitativa, textos robóticos

Abstract

This paper examines the discriminative capacity of a set of linguistic descriptors and statistical language laws (Zipf's law, Heaps' law, and the brevity principle) to distinguish human-authored from machine-generated scientific articles in Spanish. Using the ROBOT-TALK corpus (Alonso Simón et al., 2024), we build a paired set of 90 articles, each with its human-authored version and six machine-generated equivalents (Bard, Claude 3.5 Sonnet, DeepSeek-V3, GPT-3.5-turbo, GPT-4, Mixtral-8x7B-Instruct). Fifteen stylometric features that do not depend on the text's specific vocabulary are evaluated via group-based cross-validation. Sentence-length features—associated with the syntactic dimension of the brevity principle—are the most discriminative individual feature group; the full feature set reaches F1=0.943, matching Burrows' Delta (2002) without using any specific lexical form. The false positive rate on human texts is below 4 % (FPR=0.033; SVM). The most recent models (Claude 3.5 Sonnet, DeepSeek-V3, GPT-4) are statistically closest to the human profile, whereas telling the LLMs apart from one another is more challenging (macro-F1=0.699). These results suggest that sentence-level linguistic descriptors can serve as interpretable, vocabulary-independent stylometric markers for authorship analysis in the LLM era.

Keywords

authorship analysis, stylometry, statistical laws of language, quantitative linguistics, machine-generated texts

1. Introducción

En los últimos años, los denominados *grandes modelos de lenguaje* (LLMs, *Large Language Models*) han transformado por completo la producción de texto escrito, desde la comunicación cotidiana hasta la escritura académica, periodística e institucional (Abburi et al., 2023). Basados en arquitecturas *Transformer* y entrenados sobre colecciones masivas de texto (Brown et al., 2020), estos son capaces de generar textos extensos, coherentes y estilísticamente plausibles (Alonso Simón et al., 2024). A diferencia de los modelos probabilísticos clásicos, que estimaban la continuación más probable de una secuencia a partir de estadísticas de co-ocurrencia (Shannon, 1948), los LLMs reproducen con notable fidelidad el registro, el estilo y el vocabulario de géneros textuales muy diversos debido a una arquitectura que les permite modelar dependencias contextuales a larga distancia sobre corpus de entrenamiento masivos (Bel Rafecas, 2025).

Es esta capacidad de emulación la que ha situado en el centro del debate a los denominados *textos robóticos* (TR o MGT, del inglés, *Machine-Generated Texts*) (Bozza et al., 2023), cuya identificación y diferenciación frente a los textos de autoría humana se ha convertido en un desafío tanto para usuarios como para lingüistas (Gehrmann et al., 2019; Zanotto & Aroyehun, 2024). El uso de LLMs en la escritura abarca desde tareas de revisión y corrección de estilo (por ejemplo, para mejorar la redacción de autores no nativos) hasta la generación *ex novo* de un texto completo; es sobre todo en este segundo extremo donde se plantean con más fuerza los interrogantes sobre plagio, desinformación y otras prácticas fraudulentas. La frontera entre ambos escenarios no siempre es nítida (y su detección plantea, de hecho, retos metodológicos similares), pero es la generación íntegra por IA la que compromete más directamente la integridad de la autoría, con especial relevancia en el ámbito académico (Abburi et al., 2023; Alonso Simón et al., 2024). La cuestión de si detrás de un texto hay un agente humano o un proceso automatizado —es decir, un *test de Turing* (1950) aplicado a los TR— presenta implicaciones directas en ciberseguridad y en la garantía de información veraz, en un contexto en el que marcos normativos recientes (Gobierno de España, 2021) empiezan a regular la transparencia de este tipo de producciones. La detección de TR se ha convertido, por consiguiente, en una línea de investigación reciente tanto en lingüística computacional como en lingüística forense, aportando esta última un marco metodológico orientado no solo a la identificación de patrones estilísticos sino también a la evaluación de su valor como evidencia lingüística (Bernal-Ortiz et al., 2026).

La lingüística cuantitativa ofrece en este contexto una perspectiva especialmente valiosa, puesto que las regularidades estadísticas que vertebran la producción lingüística humana (p. ej., la distribución de frecuencias léxicas, el crecimiento del vocabulario o la covariación entre longitud y frecuencia de palabra) no son arbitrarias, sino que responden a leyes estadísticas relativamente estables (Zipf, 1949; Heaps, 1978; Mandelbrot, 1953). Ahora bien, aunque los LLMs se entrenan sobre vastas colecciones de texto humano, ello no garantiza que reproduzcan estas leyes con fidelidad (Kudryavtseva & Kovalevskii, 2025).

La hipótesis de este trabajo sostiene que los textos robóticos presentan un ajuste diferenciado respecto a estas leyes en comparación con los textos de autoría humana, y que un análisis de sus propiedades distribucionales puede resultar determinante a la hora de atribuir autoría. El presente trabajo analiza hasta qué punto los textos producidos por LLMs se desvían de estas regularidades y si dichas desviaciones constituyen huellas estilométricas suficientes para la atribución de autoría, aplicando métricas cuantitativas derivadas de las leyes estadísticas del lenguaje. Como objetivos específicos, planteamos:

1. Analizar cuantitativamente las propiedades léxicas y sintácticas de textos humanos (TH) y textos robóticos (TR) mediante variables estilométricas derivadas de las leyes de Zipf, Heaps y el principio de brevedad (Zipf, 1949; Heaps, 1978; Mandelbrot, 1953), con el fin de describir su potencial valor como marcadores estilométricos.
2. Comparar sistemáticamente los perfiles cuantitativos TH y TR para identificar desviaciones y patrones diferenciados por ley estadística, y determinar la proximidad de cada LLM al perfil estilométrico humano.
3. Evaluar el poder discriminativo de las métricas estilométricas mediante pruebas estadísticas, un estudio de ablación por grupos de rasgos y modelos de clasificación supervisada.

El resto del trabajo se organiza de la siguiente manera: la sección 2 enmarca el estudio en la atribución de autoría y la estilometría, revisa los antecedentes directos sobre leyes estadísticas en textos generados y los enfoques computacionales para la detección de TR, y sitúa ambas líneas en el contexto del español; la sección 3 describe el corpus de trabajo, las variables estilométricas y el protocolo de evaluación; la sección 4 presenta los resultados del análisis descriptivo y la clasificación automática, y, por último, las secciones 5 y 6 discuten las implicaciones y extraen las conclusiones principales, respectivamente.

2. Trabajos relacionados

El análisis de autoría parte de la premisa de que el idiolecto deja trazas lingüísticas identificables y cuantificables (Moyano-Moreno, 2023, 2025), una cuestión que la lingüística forense evalúa en términos de fuerza de la evidencia (*likelihood ratios*, LR) más que como una simple etiqueta binaria (Ishihara, 2021; Crespo-Miguel, 2023). La estilometría traduce estas trazas en variables susceptibles de medición (Eder et al., 2016; Savoy, 2020; Wales, 2014; Langlois-Berthelot, 2021).

La lingüística cuantitativa aporta a este marco leyes estadísticas capaces de describir regularidades universales y relativamente estables en la producción lingüística humana (Zipf, 1949; Heaps, 1978; Mandelbrot, 1953), cuyo interés para el análisis de autoría reside en que no se aplican al vocabulario concreto de un texto, sino a sus propiedades distribucionales. Se trata, no obstante, de regularidades estadísticas y no de universales exentos de variación: su magnitud y su valor discriminativo dependen de la lengua y del género textual analizados, una cuestión que retomamos más adelante a propósito del español, y en la discusión de resultados (§5). Revisamos a continuación los antecedentes que han aplicado estas leyes al estudio del texto generado por LLMs (§2.1) y los enfoques computacionales desarrollados específicamente para la detección de TR (§2.2).

2.1. Leyes estadísticas del lenguaje en los TR

Las leyes estadísticas del lenguaje han comenzado a aplicarse específicamente al estudio del texto generado por LLMs, si bien de forma todavía dispersa y con resultados que dependen del modelo, la lengua y el género analizados. Estas leyes no constituyen, en sí mismas, una prueba de autoría humana: su presencia en un texto generado es, de hecho, esperable, en la medida en que los LLMs se entrenan sobre grandes colecciones de lengua humana. La cuestión es, por tanto, si el grado de ajuste y los parámetros de estas leyes difieren sistemáticamente entre textos humanos y robóticos.

El antecedente más directo compara la distribución de frecuencias léxicas observada con una distribución zipfiana (junto con rasgos de complejidad oracional y de consistencia) para distinguir texto humano de texto generado por ordenador, con una exactitud del 98 % sobre un corpus de libros (Nguyen-Son et al., 2017); se trata, no obstante, de un resultado propio de ese corpus y de ese escenario de generación/traducción, que no debe interpretarse como una garantía general para los LLMs actuales. Trabajos posteriores matizan esta conclusión desde distintos enfoques: sobre texto generado por LSTM y GPT-2, tanto la ley de Zipf como la de Heaps se cumplen, pero sus parámetros varían con la temperatura de muestreo, de modo que el mero cumplimiento de una ley no basta como criterio de detección (el rasgo relevante es más bien el grado y el patrón de la desviación, condicionado por el modelo, el muestreo, el género y la longitud del texto) (Sumedha & Rohilla, 2021). En la misma línea, la ley de Heaps y sus funciones asociadas conservan relevancia lingüística en corpus etiquetados, sentando una línea base para su uso como indicador cuantitativo (Chacoma & Zanette, 2020); y sobre resúmenes biomédicos generados con GPT-Neo a partir de inicios reales de PubMed, la aproximación al patrón humano de la ley de Heaps aumenta con el tamaño del modelo generador (Lai et al., 2023). Una adherencia igualmente estrecha a la ley de Zipf se documenta en lenguas artificiales generadas por ChatGPT, aunque sin comparación directa con producciones humanas en una lengua natural (Diamond, 2023). Por último, se han hallado diferencias sistemáticas en la distribución de frecuencias léxicas entre textos humanos y generados por IA (Kudryavtseva & Kovalevskii, 2025), así como resultados positivos al clasificar resúmenes humanos y de IA a partir de rasgos estadísticos del texto (Mathews et al., 2024).

Estos antecedentes comparten dos limitaciones que motivan directamente el presente trabajo. En primer lugar, la mayoría se centra en una única ley evaluada de manera aislada, sin combinarla con otras leyes ni con una tarea de clasificación supervisada que permita cuantificar su poder discriminativo. En segundo lugar, ninguno se ha llevado a cabo sobre textos en español, y los géneros analizados distan del artículo científico completo en el ámbito humanístico que aquí se aborda. Advierten, además, contra un criterio dicotómico de cumplimiento o incumplimiento de una ley: el ajuste puede ser estrecho en ambos tipos de autoría y, aun así, sus parámetros diferir sistemáticamente entre ellas, como muestran los resultados de este estudio (§4).

2.2. Enfoques computacionales para la detección de TR

La detección de TR se ha formulado habitualmente como una tarea de clasificación binaria: TR frente a TH (Alonso Simón et al., 2023, 2024). Estos sistemas pueden agruparse, a grandes rasgos, en función de si construyen rasgos de superficie sobre los que aplicar aprendizaje automático clásico, explotan probabilidades internas del propio modelo generador o aprenden representaciones mediante aprendizaje profundo, una taxonomía actualizada para el escenario posterior a los LLMs conversacionales (Jawahar et al., 2020; Huang et al., 2025). Un primer grupo de sistemas emplea clasificadores como SVM, Random Forest o Naïve Bayes sobre rasgos de superficie (n-gramas de carácter y de *token*, etiquetas gramaticales, palabras funcionales o medidas de diversidad léxica) (Pizarro, 2019; Eder et al., 2016). En

esta línea, los n-gramas de orden superior ($n > 2$) resultan significativamente más frecuentes en textos generados por IA (Abburi et al., 2023), patrón validado en tareas compartidas como *SemEval-2024 Task 8* y *AuTexTification 2023* (Sharma & Mansuri, 2024; Alonso Simón et al., 2023), si bien estas cifras de rendimiento no son necesariamente extrapolables a otras lenguas o géneros textuales (Berber Sardinha, 2024). En el ámbito específicamente forense-computacional, Savoy (2020) ofrece un marco de referencia directamente pertinente para la comparación de clasificadores realizada en este trabajo (§3.3), al sistematizar los algoritmos de clasificación (desde vecinos más cercanos hasta SVM y redes neuronales) y los vectores de rasgos habituales en atribución de autoría y perfilado (palabras funcionales, n-gramas de caracteres, distancias como Delta; Burrows, 2002). Dentro de este marco de los *likelihood ratios*, un enfoque bayesiano cuantifica la redundancia estadística de los n-gramas para un análisis probabilístico del origen del texto (Bozza et al., 2023); sobre el corpus ROBOT-TALK, el mismo empleado en este trabajo, este marco probabilístico se aplica combinando marcadores discursivos, n-gramas POS y hápax (Bernal-Ortiz et al., 2026).

Junto a estos enfoques, han surgido detectores basados en propiedades internas de los propios modelos de lenguaje: *GLTR* (Gehrmann et al., 2019) visualiza la distribución de probabilidades de *tokens* para estimar si un texto pudo haber sido generado por un LLM, mientras que *DetectGPT* (Mitchell et al., 2023) explota, en cambio, la sensibilidad probabilística del modelo a perturbaciones del texto para discriminar sin necesidad de entrenamiento supervisado. Estos métodos, aunque alcanzan altas tasas de detección en condiciones controladas, requieren acceso a las distribuciones internas del modelo generativo, limitando su aplicabilidad forense cuando dicho acceso no está disponible o se desconoce el modelo que produjo el texto. Una línea distinta, no siempre recogida en las revisiones estilométricas, entrena representaciones neuronales (*embeddings* de autoría, ajuste fino de modelos preentrenados) directamente sobre la tarea de clasificación: estos sistemas suelen superar a los enfoques puramente estilométricos en condiciones de entrenamiento y prueba homogéneas, pero a costa de una menor interpretabilidad directa de los rasgos y de una mayor sensibilidad a cambios de dominio, modelo o estrategia de generación (Huang et al., 2025); en escenarios de coescritura humano-IA, por ejemplo, estos *embeddings* de autoría superan a los detectores previos en la tarea de detección binaria, pero su rendimiento se degrada en tareas más finas de verificación y atribución de autoría, donde los modelos basados en n-gramas mantienen mejor su rendimiento (Richburg et al., 2024) (véase también §5).

La evidencia empírica sobre estos enfoques en español es, sin embargo, todavía escasa: el corpus ROBOT-TALK, empleado en este trabajo, y los análisis previos realizados sobre él (Alonso Simón et al., 2024; Bernal-Ortiz et al., 2026) son, hasta donde hemos podido comprobar, los únicos antecedentes que comparan sistemáticamente TH y TR en español, y ninguno evalúa conjuntamente las leyes de Zipf, Heaps y el principio de brevedad como marcadores de clasificación. Además, ni los parámetros de estas leyes ni el rendimiento de los clasificadores son necesariamente estables entre lenguas, géneros o modelos generativos (Berber Sardinha, 2024).

3. Metodología

La metodología propuesta en este trabajo se articula como un experimento multiautor, pareado y balanceado, en tanto que se estructura en tres fases sucesivas: la primera (§3.1) describe la constitución del corpus experimental a partir de ROBOT-TALK (Alonso Simón et al., 2024), así como el procedimiento de muestreo aplicado para garantizar la comparabilidad entre textos; la segunda (§3.2) presenta el conjunto de variables estilométricas extraídas, fundamentadas en las leyes clásicas de la lingüística cuantitativa y diseñadas para capturar propiedades distribucionales que no dependen del vocabulario concreto del texto, mientras que la tercera (§3.3) detalla el análisis estadístico y el procedimiento de clasificación empleados para discernir TR de TH.

3.1. Corpus

El corpus de trabajo se basa en ROBOT-TALK (Alonso Simón et al., 2024)[1], una colección monolingüe de textos escritos en español, procedentes de diversas fuentes y géneros. Para controlar el género discursivo, el análisis se restringe a artículos científicos del ámbito de la lingüística: se trata de un registro altamente convencionalizado, en el que los LLMs suelen desenvolverse con especial competencia, lo que lo convierte en un caso de prueba más exigente para la detección que géneros con mayor variación idiosincrásica (Alonso Simón et al., 2024).

[1] Los parámetros de generación de los TR no están bajo el control de este estudio. La documentación completa al respecto figura en el corpus original (Alonso Simón et al., 2024).

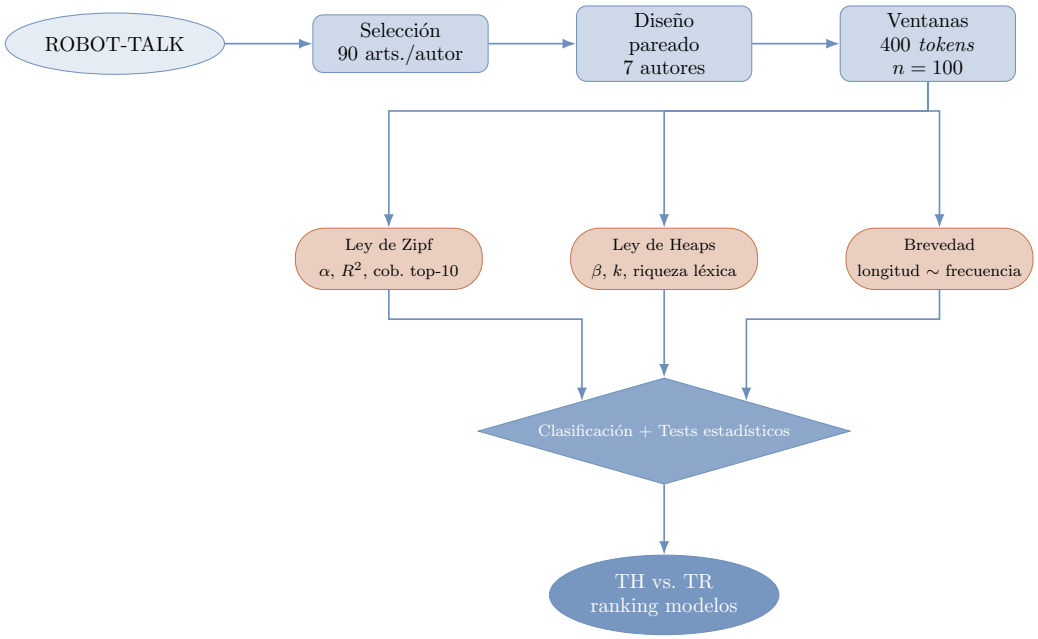

Figura 1: Flujo de la metodología. A partir del corpus ROBOT-TALK se construye un conjunto pareado de 90 artículos por autor; de cada artículo se extraen 100 ventanas aleatorias de 400 *tokens*, para aumentar la densidad de muestras, y se calculan 15 rasgos estilométricos agrupados según las tres leyes estadísticas del lenguaje; los rasgos se evalúan mediante clasificación supervisada y pruebas estadísticas para obtener el ranking de proximidad al perfil humano y las métricas de discriminación TH/TR.

Para este trabajo, se seleccionan 90 artículos científicos que disponen de seis versiones robóticas comunes: Bard, Claude 3.5 Sonnet, DeepSeek-V3, GPT-3.5-turbo, GPT-4 y Mixtral-8x7B-Instruct. Así, el corpus experimental contiene 630 textos: 90 humanos y 540 robóticos, con 90 textos por cada uno de los siete autores, tal y como se recoge en el Cuadro 1, manteniendo así la comparabilidad por artículo base y evitando que las diferencias observadas se deban a un número desigual de textos por autor.

Dado que la longitud de los documentos puede variar considerablemente entre autores, se aplica una estrategia de muestreo aleatorio de ventanas contiguas. Cada texto se segmenta en 100 muestras de longitud fija (400 *tokens*), generadas mediante muestreo aleatorio de posición inicial con reemplazo entre ventanas, sobre las que se calculan las métricas estilométricas. Se estableció $n = 100$ como número de iteraciones, suficiente para obtener estimaciones estables en las condiciones del estudio. El tamaño de ventana de 400 *tokens* es suficiente para la estimación estable de los parámetros de Zipf y Heaps, y todos los textos del corpus superan este umbral (longitud mínima observada: 409 *tokens*; Cuadro 1). Un análisis de sensibilidad con ventanas de 200, 400, 600 y 800 *tokens* confirma que los resultados son sólidos para este parámetro: la F1 varía menos de 0,015 puntos entre condiciones ($F1 \in [0{,}935, 0{,}947]$; Cuadro 8), con una ligera reducción de cobertura para $W = 800$ (540 de 630 textos) debida a documentos con longitud insuficiente. Las iteraciones de cada documento se agregan mediante la mediana para obtener una observación por texto.

3.2. Variables estilométricas

A partir de cada ventana de 400 *tokens*, se extrae un conjunto de variables estilométricas destinadas a capturar propiedades estadísticas fundamentales del lenguaje natural. Estas variables se basan en tres leyes clásicas de la lingüística cuantitativa (Zipf, 1949; Heaps, 1978; Mandelbrot, 1953), ilustradas sobre el corpus humano en la Figura 2. La elección de estas leyes responde a su carácter universal y su potencial para revelar patrones diferenciales en la producción lingüística humana frente a la artificial (Chacoma & Zanette, 2020; Diamond, 2023):

1. **Ley de Zipf** (1949). En cualquier corpus suficientemente extenso, unas pocas palabras concentran la mayor parte de los usos mientras que la inmensa mayoría aparece esporádicamente: la frecuencia f cae como función potencial del rango r:

$$\log f(r) = c + \alpha \log r \qquad (\alpha < 0)$$

		Tokens por texto				Riqueza léxica		Longitudes medias	
Autor	**Tipo**	**Mediana**	σ	**Mínimo**	**Máximo**	**TTR**	**Hápax**	**Longitud media oración**	**Longitud media palabra**
Humano	TH	1 197	457	508	2 763	0,379	0,247	33,9	5,39
Bard	TR	969	234	409	1 703	0,305	0,168	26,9	5,59
Claude 3.5 Sonnet	TR	981	200	480	1 743	0,386	0,249	27,1	6,12
DeepSeek-V3	TR	1 275	345	542	2 096	0,350	0,229	31,0	5,71
GPT-3.5-turbo	TR	1 044	257	495	1 530	0,317	0,188	29,4	5,67
GPT-4	TR	983	140	576	1 321	0,373	0,250	30,9	5,72
Mixtral-8x7B	TR	1 119	284	562	2 456	0,472	0,371	30,5	5,92
Total		1 045	314	409	2 763	0,366	0,239	29,8	5,73

Cuadro 1: Estadísticas descriptivas del corpus experimental (medianas sobre el texto completo; 90 textos por autor). σ = desviación estándar; *TTR* = *type-token ratio*; *Hápax* = proporción de palabras que aparecen una sola vez. TH = texto humano; TR = texto robótico.

donde r es el rango de la palabra ($r = 1$: la más frecuente), $f(r)$ su frecuencia de uso, $\alpha < 0$ la pendiente de la curva rango–frecuencia (estimada mediante regresión *log-log*) y c una constante de normalización.

Para operacionalizar esta ley, se estima la pendiente α junto con el coeficiente de determinación R^2, que cuantifica el ajuste de la distribución observada al modelo potencial. Un $|\alpha|$ mayor refleja una distribución más concentrada: pocas palabras dominan sobre un vocabulario menos diversificado (Lai et al., 2023; Chacoma & Zanette, 2020). Se añade la cobertura acumulada de las 10 palabras más frecuentes como estimación directa de la concentración distribucional: en una distribución zipfiana, estas palabras (fundamentalmente palabras gramaticales de alta frecuencia) concentran una proporción predecible del texto, de modo que las desviaciones respecto a ese valor reflejan el grado de ajuste a la ley sin requerir el ajuste del modelo potencial, y resulta más estable que α en muestras de tamaño moderado.

2. **Ley de Heaps** (1978). Esta ley sostiene que, a medida que un texto crece, el vocabulario se expande de forma progresivamente más lenta: cada nuevo *token* tiene más probabilidades de ser una palabra ya vista (Egghe, 2007). El tamaño del vocabulario (es decir, los *types*, V) crece con el número de *tokens* (N) según una función potencial sublineal:

$$V \approx k \cdot N^{\beta}$$

donde N es el número de *tokens* (palabras totales), V el tamaño del vocabulario (palabras distintas), β el exponente de crecimiento léxico ($0 < \beta < 1$: el vocabulario crece, pero más lentamente que el número de palabras —cada nuevo *token* tiene mayor probabilidad de ser una repetición que al inicio del texto—) y k una constante de escala. La forma logarítmica $\log V = \log k + \beta \log N$ permite estimar ambos parámetros mediante regresión log-log.

Se estiman β, k y R^2 sobre los prefijos de longitud creciente de cada ventana. Valores más altos de β indican un crecimiento léxico más acelerado y mayor riqueza de vocabulario en relación con la extensión textual. Como estimaciones directas de riqueza y diversidad léxica que no dependen del ajuste del modelo, se incluyen la *type-token ratio* (TTR; proporción de formas distintas respecto al total de *tokens*) y la tasa de hápax —proporción de palabras que aparecen una sola vez, indicadora de la tendencia a introducir vocabulario nuevo en cada tramo de texto—.

3. **Principio de brevedad** (Zipf, 1949; Mandelbrot, 1953). Este principio parte de la observación de que longitud y frecuencia léxica no son independientes: cuanto más breve es una palabra, mayor es la probabilidad de que aparezca en el discurso. Esta covariación negativa entre longitud ℓ (en caracteres) y frecuencia de uso (f) se interpreta como expresión del principio de mínimo esfuerzo del hablante, que tiende a acortar las formas que usa con más frecuencia. Formalmente se representa de la siguiente manera:

$$\log f(\ell) = a + b\,\ell \qquad (b < 0)$$

donde ℓ es la longitud de la palabra en caracteres, $f(\ell)$ la frecuencia media de las palabras de esa longitud, b la pendiente de la recta de regresión (negativa: cuanto más larga la palabra, menos frecuente) y a una constante.

Para esta ley, se operacionalizan dos aspectos: por un lado, cómo varía la longitud de las palabras según su frecuencia, y, por otro, cómo se distribuyen las longitudes oracionales en los textos.

Ley / principio	Parámetros	Métricas derivadas
Zipf	α, R^2	Cobertura top-10
Heaps	β, k, R^2	TTR, tasa hápax
Brevedad léxica	pend., corr.	Long. media pal., ratio pal. largas, ratio pal. cortas
Longitud oracional	—	Long. media oración, desv. est. oración

Cuadro 2: Las 15 variables estilométricas organizadas por fundamento teórico. *Parámetros*: estimados mediante regresión *log-log* o *log*-lineal sobre la distribución de la ley correspondiente. *Métricas derivadas*: calculadas directamente sobre los recuentos del texto, sin ajuste de modelo. *Cobertura top-10* = fracción de *tokens* cubiertos por las 10 formas más frecuentes de la ventana; *TTR* = tipos/*tokens*; *tasa hápax* = proporción de tipos que aparecen exactamente una vez; *ratio pal. largas* = proporción de formas con >6 caracteres; *ratio pal. cortas* = proporción de formas con ≤3 caracteres.

Nivel léxico (longitud de palabra). Para capturar cómo se aplica este principio en las palabras individuales, estimamos la relación entre longitud y frecuencia mediante regresión *log*-lineal. Una relación más pronunciada indica que palabras frecuentes son significativamente más cortas que las infrecuentes, reflejando el principio de economía lingüística. Además calculamos la longitud media de palabra, el ratio de palabras largas (> 6 caracteres) y cortas (≤ 3 caracteres) (Ferrer-i Cancho & Moscoso del Prado Martín, 2011).

Nivel sintáctico (longitud oracional). Extendiendo esto al nivel oracional, la operacionalizamos mediante la longitud media de oración y su desviación estándar, que capturan respectivamente la elaboración sintáctica típica y la heterogeneidad estructural del texto, particularmente relevante en prosa académica. Los límites de oración se detectan automáticamente con el modelo neuronal de spaCy para español (Honnibal et al., 2020). A diferencia del resto de métricas, que se calculan sobre ventanas de muestreo y se agregan por mediana, la longitud oracional se estima sobre el texto completo para preservar la integridad de los límites oracionales.

En total, el sistema extrae 15 variables estilométricas que no dependen del vocabulario concreto del texto, organizadas en el Cuadro 2 e ilustradas en la Figura 3.

3.3. Análisis estadístico y clasificación

Para la comparación global entre autoría humana y autoría robótica agregada, se aplica la prueba de Mann-Whitney U sobre las medianas de las iteraciones de cada métrica por texto. Se elige esta prueba no paramétrica porque varias métricas (tasa de hápax, TTR, ratios de longitud) presentan distribuciones asimétricas que no cumplen el supuesto de normalidad del test t. Para el control de comparaciones múltiples se utiliza la corrección de Benjamini-Hochberg, que regula la tasa de descubrimientos falsos (FDR) y resulta más potente que la corrección de Bonferroni cuando las hipótesis son parcialmente correlacionadas, como ocurre entre métricas derivadas de una misma ley (nivel de significación: $p < 0{,}05$); la magnitud de las diferencias se cuantifica mediante la d de Cohen (Cohen, 1988), con umbrales $d < 0{,}2$ (despreciable), 0,2–0,5 (pequeño), 0,5–0,8 (medio) y $d > 0{,}8$ (grande). Dado que los 90 artículos base están presentes en todos los autores seleccionados, se realiza además un análisis pareado mediante la prueba de Wilcoxon de rangos con signo. En este análisis, cada artículo humano se compara con la mediana de sus seis versiones robóticas, conservando el control por artículo base.

Para evaluar el poder discriminativo de las métricas, se plantean dos tareas supervisadas:

- **T1 (clasificación binaria):** TH frente a TR, con los seis modelos agrupados como autoría robótica.
- **T2 (clasificación multiclase):** identificación del autor entre siete clases (humano + seis modelos).

En ambos casos se entrenan tres clasificadores supervisados, regresión logística (LR), máquina de vectores de soporte con núcleo RBF (SVM) y Random Forest (RF), con evaluación mediante validación cruzada agrupada por artículo base (*GroupKFold*, $k = 5$), impidiendo que versiones distintas del mismo artículo base aparezcan simultáneamente en entrenamiento y prueba, evitando una posible fuga de información temática. Las 100 iteraciones de cada documento se agregan mediante la mediana *antes* de cualquier partición, de modo que la unidad estadística en todos los análisis es el texto completo (630 observaciones en T1, agrupadas en 90 artículos base para el GroupKFold). El desequilibrio de clases en la T1 (90 TH frente a 540 TR, proporción 1:6) se gestiona mediante pesos de clase inversamente proporcionales a su frecuencia en LR y SVM, penalizando más los errores en la cla-

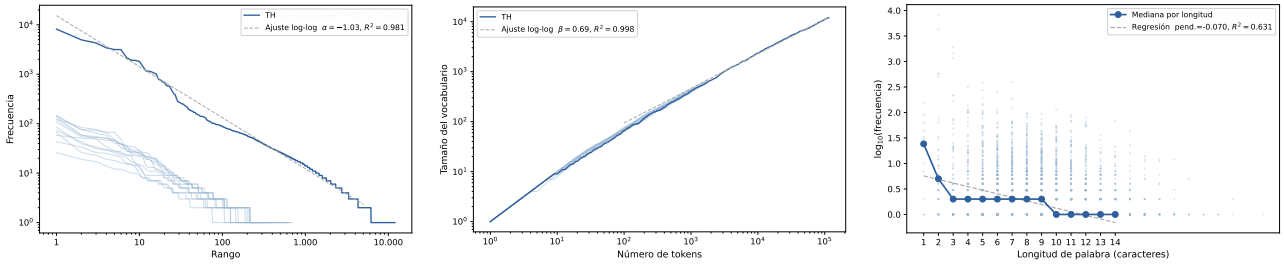

Figura 2: Las tres leyes estadísticas ilustradas sobre el corpus humano (90 artículos científicos en español). De izquierda a derecha: ley de Zipf (Zipf, 1949) — relación rango-frecuencia en escala *log-log*; las líneas azules claras corresponden a textos individuales y la línea azul al corpus agregado; ley de Heaps (Heaps, 1978) — crecimiento del vocabulario en función del número de *tokens*, y principio de brevedad (Mandelbrot, 1953) — las palabras más cortas son sistemáticamente más frecuentes (escala semilog). En los tres casos se muestra el ajuste *log-log* o *log*-lineal y sus parámetros.

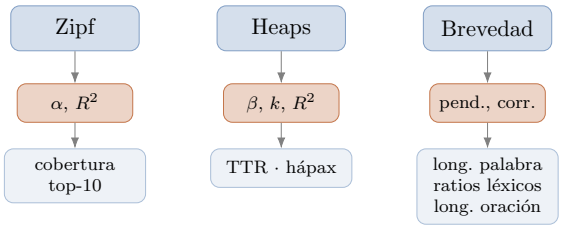

Figura 3: Organización de las 15 variables estilométricas: parámetros clásicos (naranja) y métricas derivadas independientes del vocabulario (azul).

se minoritaria (TH). Se consigna la F1 macro, AUC en T1 y la tasa de falsos positivos sobre TH ($\mathrm{FPR}_{\mathrm{TH}\to\mathrm{TR}}$ = fracción de TH clasificados como TR), especialmente relevante para su uso forense; todos estos valores se obtienen del mismo protocolo GroupKFold($k = 5$). Para cuantificar la contribución individual de cada grupo de variables, se lleva a cabo un estudio de ablación organizado según la ley de referencia de cada grupo (Figura 7 y Cuadro 9). Por último, para ordenar los modelos según su proximidad al perfil estadístico humano, se calcula una distancia global D sobre los rasgos de las tres leyes (α y R^2 de Zipf; β, k y R^2 de Heaps; pendiente y correlación de brevedad; longitud media de oración; longitud media de palabra): para cada rasgo i se obtiene la desviación relativa absoluta del modelo respecto al humano ($|m_i - h_i| \,/\, |h_i|$) y se promedian los nueve valores con igual peso.

Para garantizar la trazabilidad del procedimiento, el muestreo aleatorio de ventanas y las particiones de validación cruzada se ejecutan con una semilla aleatoria fija (`seed`=42); la segmentación oracional se realiza con el modelo `es_core_news_sm` (v. 3.8.0) de spaCy (v. 3.8.11) (Honnibal et al., 2020), y los clasificadores se implementan con scikit-learn (v. 1.8.0) sobre Python 3.11.

4. Resultados

4.1. ¿Cuál es el modelo estadísticamente más humano?

Las diferencias globales entre TH y TR quedan ilustradas en las Figuras 4 y 5. El análisis por ley revela que la ley de Zipf (1949) y la de Heaps (1978) muestran desviaciones estadísticamente significativas ($p < 0{,}05$, corrección BH) pero con tamaños de efecto pequeños (Cohen, 1988); en cambio, el principio de brevedad (Zipf, 1949; Mandelbrot, 1953) produce la separación más consistente, con efectos grandes tanto en las variables sintácticas —desviación estándar de la oración ($d = 2{,}38$) y longitud media de oración ($d = 1{,}07$)— como en las léxicas —ratio de palabras largas ($d = 1{,}89$) y longitud media de palabra ($d = 1{,}70$)— (Cohen, 1988) (Cuadro 4). El Cuadro 5 muestra la distancia relativa de cada modelo al perfil estadístico humano; la Figura 6 desglosa esa distancia por dimensión estadística.

El análisis pareado por artículo base (prueba de Wilcoxon; Cuadro 3) confirma este patrón y, en varias métricas, lo refuerza: al controlar por artículo base se elimina la variabilidad temática entre textos, y los tamaños de efecto aumentan de forma notable respecto a la comparación no pareada, especialmente en la pendiente de la brevedad léxica, que pasa de un efecto medio ($d = 0{,}64$) a uno grande ($d = 0{,}83$).

El ranking global sitúa a Claude 3.5 Sonnet, DeepSeek-V3 y GPT-4 como los modelos más próximos al perfil humano (distancias 0,043–0,046), con pendientes de Zipf ($\alpha \approx -0{,}64$–$-0{,}65$) y exponentes de Heaps ($\beta \approx 0{,}82$) prácticamente idénticos a los humanos. GPT-3.5-turbo y Bard presentan una distribución de Zipf más concentrada (α más negativo), indicando un vocabulario menos diversificado. Mixtral-8x7B ocupa el extremo opuesto:

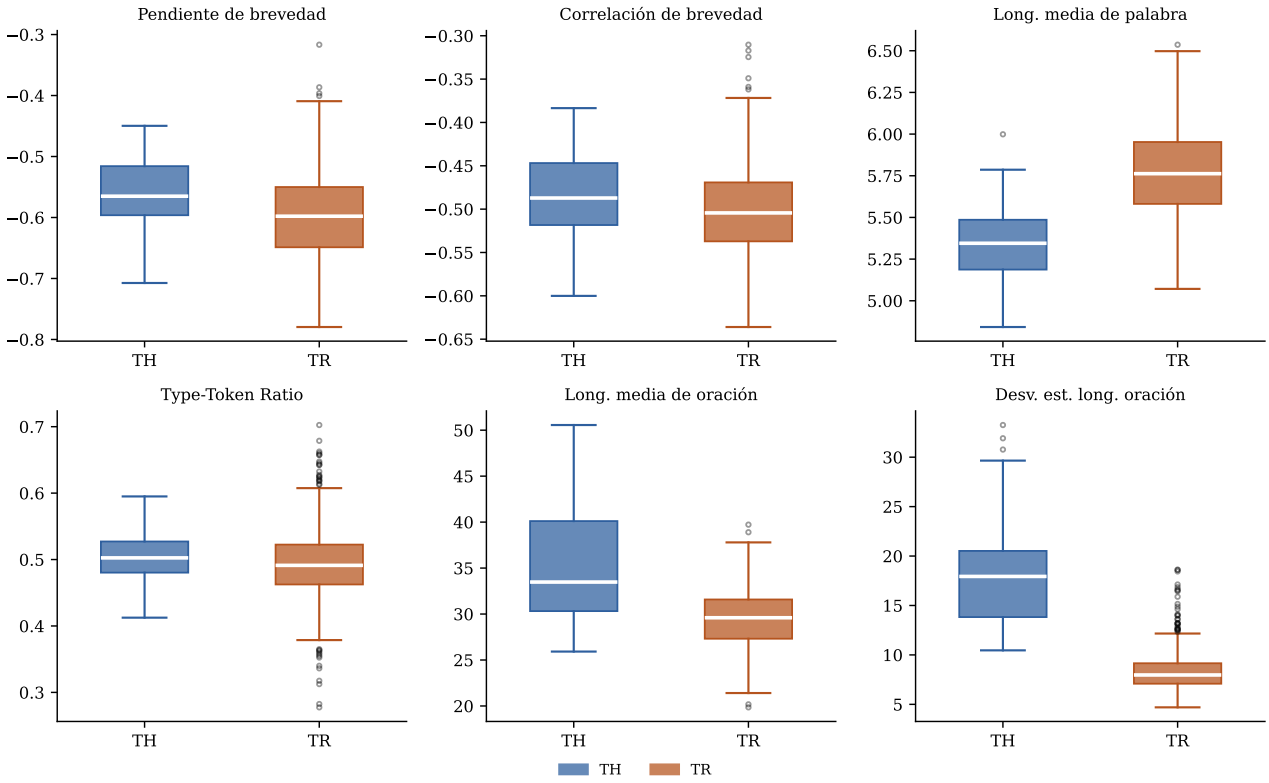

Figura 4: Distribución de las seis métricas más discriminativas por tamaño de efecto $|d|$ (Cuadro 4): desviación estándar de oración ($d = 2{,}38$), ratio palabras largas ($d = 1{,}89$), longitud media de palabra ($d = 1{,}70$), ratio palabras cortas ($d = 1{,}16$), longitud media de oración ($d = 1{,}07$) y parámetro k de Heaps ($d = 0{,}54$). TH en azul, TR en naranja; línea central = mediana; caja = IQR. La extensión del principio de brevedad (Zipf, 1949; Mandelbrot, 1953) al nivel sintáctico (longitud de oración) muestra la mayor separación entre corpus.

Variable	Med. dif.	W	$\lvert d\rvert$	Sig.
Zipf α	0,015	1248	0,36	+++
Zipf R^2	−0,010	1044	0,47	+++
Heaps β	0,008	928	0,46	+++
Heaps R^2	0,001	1218	0,16	+++
Brevedad (pend.)	0,042	503	0,83	+++
Brevedad (corr.)	0,022	986	0,52	+++

Cuadro 3: Análisis pareado por artículo base (prueba de Wilcoxon de rangos con signo; $n = 90$ pares, cada TH frente a la mediana de sus seis versiones robóticas). Med. dif. = mediana de las diferencias TH−TR; $|d| = d$ de Cohen (Cohen, 1988); Sig.: +++ $p_{\text{BH}} < 0{,}001$. Los resultados confirman el patrón del Cuadro 4: la pendiente de la brevedad léxica concentra el mayor tamaño de efecto, que aquí alcanza el rango grande.

con $\alpha = -0{,}53$ y $\beta = 0{,}86$, su distribución de frecuencias es la más plana del conjunto y, simultáneamente, el modelo con mayor riqueza léxica (TTR = 0,59).

El principio de brevedad (Zipf, 1949; Mandelbrot, 1953) revela un patrón que afecta a todos los modelos sin excepción: ninguno reproduce la longitud oracional humana. Los TH presentan las oraciones más largas del corpus ($\bar{L} = 33{,}9$ *tokens*), mientras que todos los modelos producen oraciones más cortas. La brecha es especialmente pronunciada en Bard ($\Delta = -7{,}1$) y Claude 3.5 Sonnet ($\Delta = -6{,}9$), siendo DeepSeek-V3 y GPT-4 los más próximos ($\Delta \approx -3$). En cuanto a la longitud de palabra, todos los modelos emplean palabras más largas que los TH, siendo Bard el más moderado ($\Delta = +0{,}21$) y Claude 3.5 Sonnet el más alejado ($\Delta = +0{,}74$).

4.2. ¿Hasta dónde llega el poder discriminativo de las leyes estadísticas?

El estudio de ablación (Figura 7 y Cuadro 9) cuantifica el poder discriminativo de cada grupo de rasgos. Los rasgos de Zipf se sitúan cerca del azar (F1=0,52), los de Heaps alcanzan un rendimiento moderado (F1=0,63) y las variables de longitud oracional producen la mayor separación

Variable	Med. TH	Med. TR	$\lvert d\rvert$	Sig.
Ley de Zipf				
α	−0,643	−0,657	0,17	+
R^2	0,883	0,891	0,02	+
Cobertura top-10	0,327	0,328	0,10	−
Ley de Heaps				
β	0,825	0,815	0,32	++
k	1,502	1,561	0,54	+++
R^2	0,996	0,995	0,27	++
Principio de brevedad				
Pendiente	−0,565	−0,598	0,64	+++
Correlación	−0,487	−0,504	0,41	+++
Longitud media palabra	5,345	5,763	1,70	+++
TTR	0,503	0,491	0,17	+
Tasa hápax	0,363	0,352	0,14	+
Ratio palabras largas	0,363	0,418	1,89	+++
Ratio palabras cortas	0,438	0,413	1,16	+++
Longitud media oración	33,93	29,49	1,07	+++
Desviación estándar oración	17,97	7,85	2,38	+++

Cuadro 4: Comparación TH vs. TR para las 15 variables estilométricas (Mann-Whitney U con corrección de Benjamini-Hochberg). Med. = mediana de iteraciones por texto; $|d| = d$ de Cohen (Cohen, 1988); Sig.: + $p < 0{,}05$; ++ $p < 0{,}01$; +++ $p < 0{,}001$; − no significativo. Los valores más altos en TH corresponden a α, β, pendiente, correlación, TTR, tasa hápax, ratio palabras cortas, longitud media de oración y desviación estándar de oración; los más altos en TR a R^2 Zipf, k, longitud media de palabra y ratio palabras largas.

Modelo	Rango	Zipf α	Heaps β	Long. pal. (Δ)	Long. or. (Δ)	TTR	Dist. global
Humano (referencia)	—	−0,64	0,82	5,35	33,9	0,50	—
Claude 3.5 Sonnet	**1**	−0,64	0,82	6,08 (+0,74)	27,1 (−6,9)	0,50	0,043
DeepSeek-V3	2	−0,65	0,82	5,68 (+0,33)	31,0 (−2,9)	0,50	0,046
GPT-4	3	−0,65	0,82	5,68 (+0,34)	30,9 (−3,0)	0,49	0,046
GPT-3.5-turbo	4	−0,70	0,80	5,65 (+0,30)	29,4 (−4,6)	0,46	0,066
Bard	5	−0,72	0,79	5,56 (+0,21)	26,9 (−7,1)	0,44	0,085
Mixtral-8x7B	**6**	−0,53	0,86	5,93 (+0,58)	30,5 (−3,5)	0,59	0,095

Cuadro 5: Ranking de proximidad al perfil estadístico humano (azul = más cercano; naranja = más alejado). Los valores de Zipf α, Heaps β y TTR son medianas sobre ventanas de 400 *tokens* (los valores de TTR sobre el texto completo figuran en el Cuadro 1); Δ indica la diferencia con respecto al humano. La distancia global D (definida en §3.3) ordena los modelos por proximidad al perfil humano.

individual (F1=0,89; AUC=0,98; RF). El conjunto completo de 15 métricas alcanza F1=0,943 y AUC=0,991 (Figura 8). La tasa de falsos positivos sobre TH ($\text{FPR}_{\text{TH}\to\text{TR}}$) es especialmente relevante para la atribución de autoría. El clasificador SVM sobre el conjunto completo registra FPR=0,033±0,027, es decir, menos del 4 % de los TH son erróneamente clasificados como robóticos; el Random Forest presenta un FPR notablemente más alto (0,167±0,035). Como *baseline* estilométrico, la distancia Delta (Burrows, 2002) sobre las 200 palabras más frecuentes (MFW), valor convencional en estilometría computacional (Eder et al., 2016), alcanza F1 comparable con clasificadores lineales (LR: 0,948; SVM: 0,950), pero rinde sensiblemente peor con Random Forest (F1=0,827 frente a 0,929), poniendo de manifiesto que los indicadores estadístico-lingüísticos codifican patrones no lineales que las frecuencias de palabras funcionales no capturan —es decir, combinaciones entre métricas (como oraciones largas junto a vocabulario más elaborado) que resultan informativas en conjunto pero no lo serían individualmente— (Cuadro 9). En la T2, el mejor clasificador (SVM) obtiene F1 macro=0,699 (Cuadros 6 y 7). El desglose

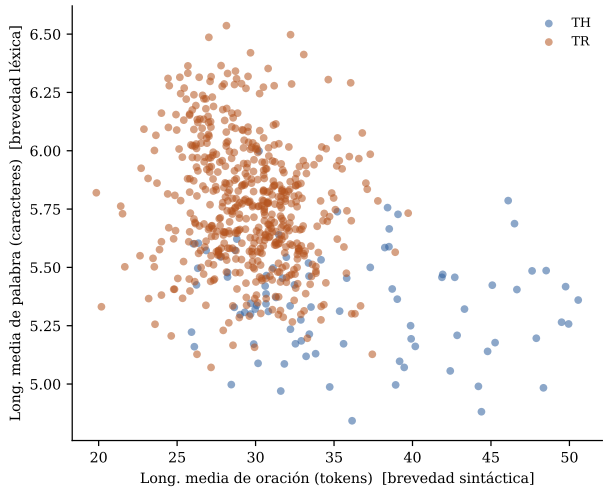

Figura 5: Dispersión TH/TR en el espacio de economía léxica (eje x: longitud media de palabra en caracteres) frente a economía sintáctica (eje y: longitud media de oración en *tokens*). Los TR emplean palabras más largas pero producen oraciones más cortas, disociando los dos niveles de economía lingüística: en los TH, mayor elaboración léxica covaría con mayor complejidad oracional; en los TR, esta asociación se rompe. Cada punto representa un texto; medianas sobre ventanas de 400 *tokens*.

Clasificador	F1 macro	Accuracy
LR	0,692 ±0,018	0,694 ±0,018
SVM (RBF)	0,699 ±0,034	0,702 ±0,028
RF	0,670 ±0,034	0,675 ±0,029

Cuadro 6: Identificación multiclase de autor (7 clases: humano + 6 modelos). Macro-F1 y Accuracy (media ± desv. est.), conjunto completo de 15 rasgos (GroupKFold, k=5). LR = Regresión logística; RF = Random Forest.

por clase revela que Mixtral-8x7B (F1=0,949) y los TH (F1=0,913) son las clases más identificables, mientras que GPT-3.5-turbo (F1=0,511) y DeepSeek-V3 (F1=0,551) presentan mayor confusión entre clases.

5. Discusión

Los resultados de este estudio se inscriben en una línea de investigaciones recientes que han evaluado la competencia lingüística de los LLMs (Mahowald et al., 2024) y, en particular, el grado en que estos reproducen las leyes estadísticas universales observadas en el lenguaje natural (Chacoma & Zanette, 2020; Diamond, 2023; Lai et al., 2023; Mathews et al., 2024). La hipótesis de partida —que los TR presentan un ajuste diferenciado

Clase	Prec.	Rec.	F1
Mixtral-8x7B	0,976	0,922	0,949
Humano	0,894	0,933	0,913
Claude 3.5 Sonnet	0,755	0,856	0,802
Bard	0,662	0,544	0,598
GPT-4	0,550	0,611	0,579
DeepSeek-V3	0,597	0,511	0,551
GPT-3.5-turbo	0,490	0,533	0,511
Macro avg.	**0,703**	**0,702**	**0,700**

Cuadro 7: Desglose por clase en la T2 (SVM, GroupKFold k=5), ordenado por F1 descendente.

W (*tokens*)	Textos	F1	AUC
200	630	0,947 ±0,023	0,989 ±0,008
400	630	0,946 ±0,026	0,990 ±0,006
600	617	0,947 ±0,020	0,996 ±0,004
800	540	0,935 ±0,047	0,993 ±0,009

Cuadro 8: Sensibilidad al tamaño de ventana de muestreo (SVM, GroupKFold, k=5, conjunto completo de 15 rasgos). La reducción de cobertura en W=800 se debe a textos con menos de 800 *tokens*.

a las leyes estadísticas del lenguaje natural (Zipf, 1949; Heaps, 1978; Mandelbrot, 1953) respecto a los TH— queda confirmada, si bien no todas las leyes son igual de discriminativas: las desviaciones en Zipf y Heaps son estadísticamente detectables pero de magnitud reducida, mientras que el principio de brevedad, especialmente en su dimensión oracional, concentra la mayor parte del poder discriminativo. Esto sugiere que el grado de cumplimiento formal de estas leyes no es, en sí mismo, lo que resulta discriminante, sino las variables lingüísticas concretas que miden (principalmente la longitud de palabra y, sobre todo, la longitud de oración).

Los LLMs se entrenan sobre grandes corpus de textos humanos y aprenden a reproducir con relativa fidelidad la distribución de frecuencias léxicas y el crecimiento del vocabulario, dimensiones capturadas por Zipf (1949) y Heaps (1978), pero no logran replicar, en este caso, los patrones de longitud oracional propios de la prosa académica humana. La longitud de oración en los TH del corpus ($\bar{L}$ = 33,9 *tokens*) supera sistemáticamente a la de todos los modelos (26,9–31,0), lo que sugiere que los LLMs tienden a generar una prosa más fragmentada incluso cuando imitan el registro académico. Esta tendencia de generar oraciones más cortas y uniformes que los humanos ha sido documentada en trabajos com-

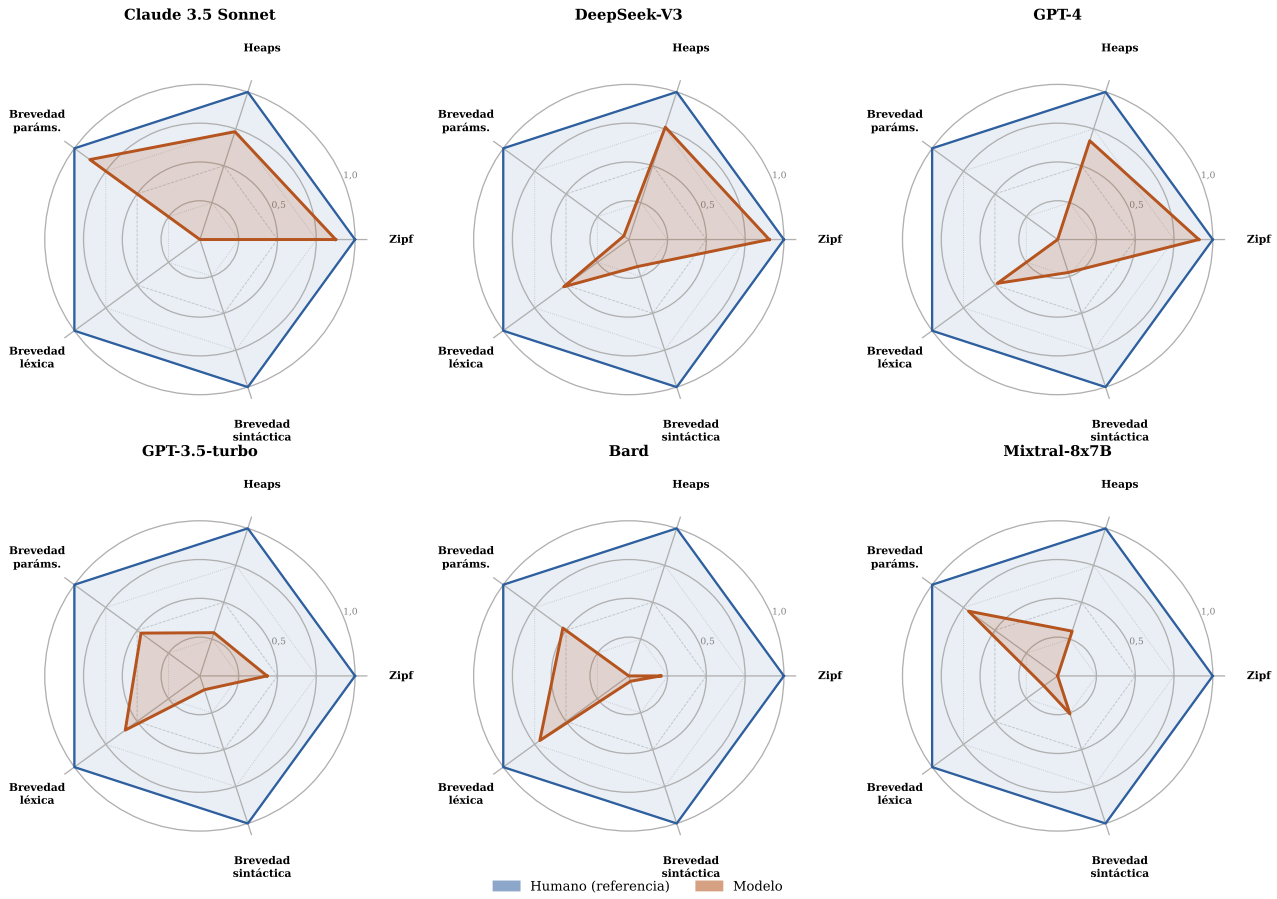

Figura 6: Proximidad al perfil estilométrico humano (azul) por dimensión para cada modelo (naranja); 1,0 = igual al humano, 0 = máxima desviación del conjunto. Los modelos recientes (Claude 3.5 Sonnet, DeepSeek-V3, GPT-4) se acercan en Zipf (1949) y Heaps (1978); la longitud oracional (Mandelbrot, 1953) es la dimensión en que todos los modelos se alejan más del perfil humano.

parables sobre otras lenguas y géneros (Zanotto & Aroyehun, 2024; Alonso Simón et al., 2024), lo que apunta a que se trata de una propiedad distribucional general de estos modelos, no de un artefacto específico del dominio académico en español. La interpretación más plausible es que los LLMs tienden a producir oraciones más cortas con un léxico más elaborado, en detrimento de la complejidad retórica y argumentativa que caracteriza, generalmente, a la escritura académica humana (Gehrmann et al., 2019; Bozza et al., 2023). Por otro lado, esta marcada uniformidad oracional observada en los TR podría estar influida por los parámetros de decodificación predeterminados (tales como la temperatura o el muestreo *top-p*) empleados en la generación de las respuestas. Dado que en este trabajo el foco principal se situó en evaluar la huella estilométrica bajo la configuración por defecto de los modelos, queda como trabajo futuro examinar de forma sistemática el impacto del ajuste explícito de la temperatura sobre la variabilidad sintáctica y la longitud oracional, así como diversificar el análisis explorando distintas modalidades y subgéneros dentro de la propia escritura académica.

Desde el punto de vista de la clasificación automática, la distancia Delta (Burrows, 2002) y el conjunto completo de rasgos estadísticos alcanzan un rendimiento similar con clasificadores lineales (F1≈0,95). Esta comparación debe interpretarse con cautela: Delta opera sobre 200 MFW frente a las 15 métricas estadístico-lingüísticas propuestas; fue diseñado para la atribución de autoría entre textos humanos, no para la detección de IA; y Random Forest no es su clasificador natural, ya que Delta presupone distancias en espacio de rasgos de frecuencia, lo que explica su rendimiento más bajo con RF. La comparación no es, por tanto, una evaluación en igualdad de condiciones, sino una referencia de orden de magnitud. Aunque el AUC del Delta con clasificadores lineales (0,995–0,996) es marginalmente superior al del enfoque propuesto (0,985–0,991), el verdadero valor del enfoque propuesto radica en que logra esta misma eficacia de forma independiente a la identidad de las palabras: los 15 rasgos

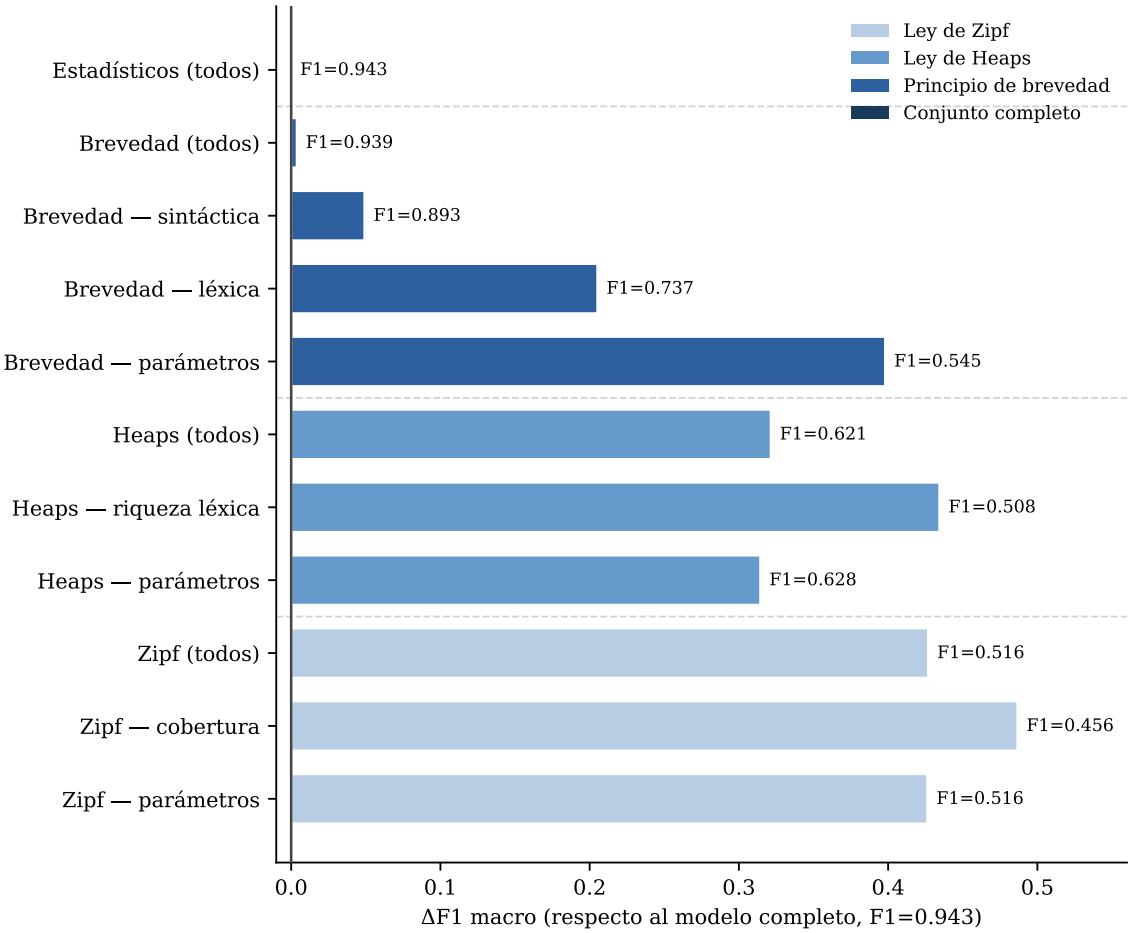

Figura 7: Estudio de ablación: ΔF1 macro de cada subconjunto de rasgos respecto al modelo completo (F1 = 0,943, SVM, línea vertical). Las barras más largas indican mayor déficit al usar ese grupo en solitario; el F1 absoluto del mejor clasificador figura a la derecha de cada barra. Los grupos se organizan por ley estadística de referencia.

no recurren a ninguna forma léxica concreta (ni palabras funcionales ni de contenido específico), lo que plantea la posibilidad de su adaptación a otras lenguas, siempre que se empleen procedimientos de tokenización y segmentación adecuados y se evalúe previamente la estabilidad de las métricas. Conviene precisar que algunos rasgos (TTR, hápax, longitudes de palabra, cobertura top-10) sí dependen de propiedades distribucionales del vocabulario realizado; la independencia reside en que no se usa la identidad de ninguna palabra concreta, no en que los rasgos sean ajenos al léxico como fenómeno. Cabe señalar, no obstante, que la segmentación oracional mediante spaCy y las medidas de longitud en caracteres introducen cierta dependencia de las convenciones ortográficas de la lengua, por lo que su extensión a lenguas con sistemas de escritura muy diferentes requiere la selección de herramientas apropiadas. Además, el Random Forest obtiene un rendimiento notablemente superior con rasgos estadísticos (F1=0,929) que con el Delta (F1=0,827), lo que indica que los primeros codifican patrones no lineales que las frecuencias de palabras funcionales no capturan —es decir, la asociación entre oraciones largas y vocabulario más elaborado resulta informativa en conjunto pero no lo sería individualmente— (Cuadro 9).

Como punto de comparación con los detectores léxicos de referencia en las principales tareas compartidas —*SemEval-2024 Task 8* (Sharma & Mansuri, 2024) y *AuTexTification 2023* (Sarvazyan et al., 2023)—, se evaluó un clasificador SVM lineal sobre vectores TF-IDF de n-gramas de carácter (orden 2–4, 50 000 rasgos), una de las representaciones más habituales en estos sistemas. Este *baseline* léxico alcanza F1=0,973±0,023 y AUC=0,999±0,001, superando en F1 al enfoque estadístico-lingüístico (Cuadro 9). Sin embargo, su tasa de falsos positivos sobre TH es mayor (FPR=0,056±0,070) que la del SVM estadístico (FPR=0,033±0,027): el enfoque propuesto es más conservador en la clasificación de TH. Además, al no depender de unida-

Grupo de rasgos	LR		SVM		RF	
	F1	AUC	F1	AUC	F1	AUC
Ley de Zipf						
Parámetros (α, R^2)	0,451±0,04	0,559±0,04	0,451±0,05	0,631±0,03	0,516±0,05	0,569±0,02
Todos	0,457±0,04	0,569±0,07	0,446±0,03	0,589±0,02	0,516±0,05	0,523±0,05
Ley de Heaps						
Parámetros (β, k, R^2)	0,597±0,05	0,739±0,05	0,628±0,03	0,778±0,03	0,552±0,04	0,711±0,04
Riqueza léxica (TTR, hápax)	0,451±0,04	0,545±0,03	0,436±0,05	0,641±0,02	0,508±0,03	0,585±0,03
Todos	0,596±0,03	0,764±0,05	0,621±0,05	0,789±0,03	0,583±0,07	0,705±0,04
Principio de brevedad						
Parámetros (pend., corr.)	0,528±0,02	0,678±0,04	0,519±0,03	0,713±0,04	0,545±0,03	0,627±0,04
Léxica (long. y ratios)	0,737±0,03	0,906±0,01	0,723±0,03	0,878±0,01	0,697±0,02	0,858±0,02
Longitud oracional	**0,878**±0,03	**0,982**±0,01	**0,869**±0,03	**0,971**±0,01	**0,893**±0,04	**0,983**±0,01
Todos	0,907±0,04	0,987±0,01	0,928±0,03	0,984±0,02	0,939±0,03	0,987±0,01
Estadísticos (todos, n=15)	**0,926**±0,03	**0,985**±0,01	**0,943**±0,03	**0,991**±0,01	**0,929**±0,01	**0,991**±0,01
Baseline estilométrico (palabras funcionales)						
Delta Burrows (top-200)	0,948±0,02	0,996±0,00	0,950±0,03	0,995±0,00	0,827±0,07	0,994±0,00
n-gramas de carácter TF-IDF†	—		0,973±0,023	0,999±0,001	—	

Cuadro 9: Ablación completa — macro-F1 y AUC (media ± desv. est.) para los tres clasificadores en la T1 (TH vs. TR; GroupKFold por artículo base, k=5). LR = Regresión logística; SVM = máquina de vectores de soporte (RBF); RF = Random Forest. El *baseline* Delta de Burrows usa frecuencias relativas z-normalizadas de las 200 palabras más frecuentes (MFW). †n-gramas de carácter: TF-IDF orden 2–4, 50 000 rasgos, evaluado únicamente con SVM lineal calibrado.

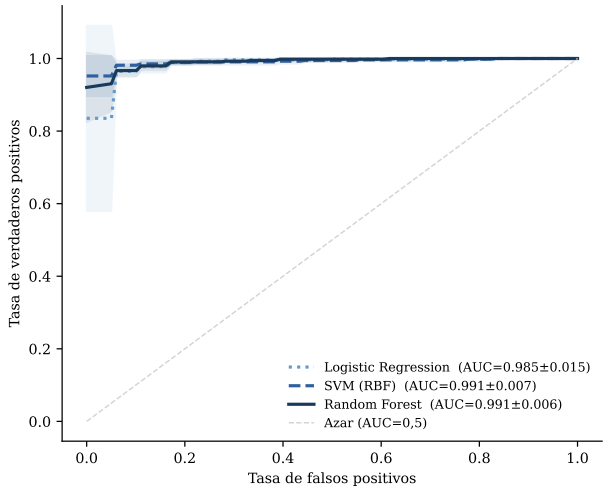

Figura 8: Curvas ROC de los tres clasificadores sobre el conjunto completo de 15 rasgos (GroupKFold, $k = 5$). La banda sombreada representa ±1 desviación estándar entre *folds*.

des léxicas específicas, el enfoque puede aplicarse a textos de distinta temática sin riesgo de contaminación léxica.

Sobre el mismo corpus ROBOT-TALK, Bernal-Ortiz et al. (2026) aplican un enfoque híbrido forense basado en rasgos identificados manualmente (marcadores discursivos, n-gramas POS y hápax) en el marco de los *likelihood ratios*, alcanzando hasta un 99,45 % de precisión computacional y cobertura forense del 100 % sobre los textos que superan el umbral LR ≥ 100 (aproximadamente el 55 % del corpus). El presente trabajo resulta complementario: con 15 rasgos extraídos de forma completamente automática se alcanza F1=0,943 con cobertura del 100 % de los textos, a expensas de una menor interpretabilidad directa sobre los marcadores lingüísticos concretos.

Los resultados de la T2 (F1 macro=0,699) muestran que las métricas propuestas distinguen con claridad TH de TR, pero captan solo parcialmente las diferencias entre modelos generativos. Esto es coherente con la hipótesis de que los LLMs comparten rasgos estilométricos comunes que los diferencian del lenguaje humano como clase, aunque sus idiolectos específicos sean más difíciles de separar con los rasgos empleados.

El alcance de las conclusiones está condicionado por el hecho de que todas las evaluaciones se han realizado sobre un único corpus, género y lengua: ni los parámetros de las leyes estadísticas del lenguaje ni el rendimiento de los clasificadores son necesariamente estables entre lenguas, géneros o modelos generativos (Berber Sardinha, 2024), por lo que la replicación en otras lenguas y géneros resulta necesaria para determinar si las desviaciones documentadas son propiedades

generales de los LLMs o si, por el contrario, se trata de particularidades del español académico. El desglose por clases en la T2 revela que los modelos no son igualmente difíciles de identificar: mientras que Mixtral-8x7B resulta muy reconocible, GPT-3.5-turbo se confunde con otros modelos, lo que apunta a la viabilidad de elaborar perfiles estilométricos específicos por familia de modelos, una dirección prometedora para tareas de atribución de autoría entre LLMs. Dos escenarios quedan pendientes de explorar: la reescritura del texto generado por IA (mediante paráfrasis humana o automática) y la manipulación deliberada de los rasgos, por ejemplo alargando artificialmente las oraciones para aproximar el texto al patrón humano, por lo que resulta fundamental abordar en trabajos futuros la fiabilidad de las métricas en escenarios de coescritura, postedición o uso de LLMs como asistentes de redacción (*proofreading*), donde la frontera entre la autoría humana, la robótica y la híbrida se difumina significativamente (Queralt et al., 2025). Finalmente, a medida que los LLMs aumenten en tamaño y complejidad, podrían aproximarse con mayor fidelidad a las distribuciones estadísticas del lenguaje humano (Diamond, 2023), lo que reforzaría la necesidad de una monitorización continua de la validez de estos marcadores.

6. Conclusión

Este trabajo ha mostrado que los descriptores lingüísticos derivados de las leyes estadísticas del lenguaje constituyen huellas estilométricas capaces de discriminar TH de textos generados por modelos de lenguaje en artículos científicos en español. Sobre un conjunto pareado y balanceado de 630 textos extraídos del corpus ROBOT-TALK, las 15 métricas alcanzan una F1 macro de 0,943 y un AUC de 0,991 en la T1, con un rendimiento comparable al de la distancia Delta (§5). La longitud oracional es el rasgo singular más discriminativo (F1=0,89; RF), mientras que la T2 resulta más exigente (F1=0,699), lo que apunta a que los modelos son más fáciles de distinguir de los TH que entre ellos mismos. Los tres objetivos del trabajo quedan así cumplidos: (1) se han caracterizado cuantitativamente las propiedades léxicas y sintácticas de TH y TR mediante las leyes estadísticas del lenguaje; (2) se han identificado las desviaciones significativas por dimensión y la jerarquía de modelos en proximidad al perfil humano, y (3) se ha evaluado el poder discriminativo de las métricas en términos de clasificación automática y tasa de falsos positivos, con atención a su relevancia para el análisis de autoría. La tasa de falsos positivos obtenida ($FPR_{TH \to TR}$=0,033; SVM) es favorable para un sistema de clasificación automática; no obstante, los marcos forenses imponen estándares de evidencia considerablemente más estrictos, por lo que estos resultados deben interpretarse como indicios estadísticos de autoría robótica y no como evidencia probatoria directa.

La estilometría aplicada a la detección de textos generados por IA se adentra, pues, en un terreno más que fértil: a medida que los LLMs aumenten en tamaño y complejidad, podrían aproximarse con mayor fidelidad a las distribuciones estadísticas del lenguaje humano, difuminando progresivamente las diferencias documentadas en este trabajo. Sin embargo, su aprovechamiento sistemático ofrece una perspectiva complementaria tanto para la lingüística cuantitativa como para la lingüística forense en un momento marcado por la generación automática de texto.

Disponibilidad de código y datos

El código, las tablas de resultados y las figuras de este trabajo están disponibles en Zenodo (https://doi.org/10.5281/zenodo.24306545) y en GitHub (https://github.com/isabel-mm/statlaws-mgt-es). El corpus ROBOT-TALK no se redistribuye; el repositorio incluye los rasgos estilométricos ya calculados, que permiten reproducir los análisis posteriores a su extracción.

Agradecimientos

Este trabajo se enmarca en el proyecto «ROBOT-TALK: Reconocimiento del origen robótico de textos. Automatización de tareas y conocimiento lingüístico» (PID2022-140897OB-I00), financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea (FEDER). Esta investigación está financiada por la ayuda predoctoral FPU21/03261 concedida por el Ministerio de Universidades.

Referencias

Abburi, Harika, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen & Sanmitra Bhattacharya. 2023. Generative AI text classification using ensemble LLM approaches. arXiv [cs.CL/cs.AI]. 10.48550/arXiv.2309.07755

Alonso Simón, Lara, Ana M. Fernández-Pampillón Cesteros, Marianela Fernández Trinidad & Manuel M. Márquez Cruz. 2024. ¿Tienen GPT-3.5 y GPT-4 un estilo de escritura diferente del estilo humano? Un estudio exploratorio para el español. *Revista Electrónica de Lingüística Aplicada* 23(1). 34–54. 10.58859/rael.v23i1.666

Alonso Simón, Lara, José Antonio Gonzalo Gimeno, Ana M. Fernández-Pampillón Cesteros, Marianela Fernández Trinidad & Victoria Escandell Vidal. 2023. Using linguistic knowledge for automated text identification. En *Iberian Languages Evaluation Forum (IberLEF)*,

Bel Rafecas, Núria. 2025. Un vocabulario básico de los modelos grandes de lenguaje: una explicación y 25 términos. *Terminàlia* 31. 27–39. 10.2436/20.2503.01.213

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

Bernal-Ortiz, Francisco, Isabel Moyano-Moreno & Mario Crespo-Miguel. 2026. Machine-generated vs. human texts: A hybrid approach to forensic feature-based text classification. *International Journal of Interactive Multimedia and Artificial Intelligence* En prensa

Bozza, Silvia, Claude-Alain Roten, Antoine Jover, Valentina Cammarota, Lionel Pousaz & Franco Taroni. 2023. A model-independent redundancy measure for human versus ChatGPT authorship discrimination using a Bayesian probabilistic approach. *Scientific Reports* 13(1). 19217. 10.1038/s41598-023-46390-8

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33. 1877–1901

Burrows, John. 2002. Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17(3). 267–287. 10.1093/llc/17.3.267

Ferrer-i Cancho, Ramon & Fermín Moscoso del Prado Martín. 2011. Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment* 2011(12). L12002. 10.1088/1742-5468/2011/12/L12002

Chacoma, Andrés & Damián H. Zanette. 2020. Heaps' Law and Heaps functions in tagged texts: evidences of their linguistic relevance. *Royal Society Open Science* 7(3). 200008. 10.1098/rsos.200008

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates 2ª ed.

Crespo-Miguel, Mario. 2023. El análisis de autoría en Lingüística forense: Historia, concepción y revisión metodológica. Aplicación de la razón de verosimilitud a textos cortos en español. *Revista Signos. Estudios de Lingüística* 56(111). 35–58. 10.4067/S0718-09342023000100035

Diamond, Justin. 2023. Genlangs and Zipf's Law: Do languages generated by ChatGPT statistically look human? arXiv [cs.CL]. 10.48550/arXiv.2304.12191

Eder, Maciej, Jan Rybicki & Mike Kestemont. 2016. Stylometry with R: a package for computational text analysis. *The R Journal* 8(1). 107–121. 10.32614/RJ-2016-007

Egghe, Leo. 2007. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments. *Journal of the American Society for Information Science and Technology* 58(5). 702–709. 10.1002/asi.20524

Gehrmann, Sebastian, Hendrik Strobelt & Alexander M. Rush. 2019. GLTR: Statistical detection and visualization of generated text. En *57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 111–116. 10.18653/v1/P19-3019

Gobierno de España. 2021. Carta de Derechos Digitales. Ministerio de Asuntos Económicos y Transformación Digital.

Heaps, Harold Stanley. 1978. *Information retrieval: Computational and theoretical aspects*. Academic Press. Section 7.5, pp. 206–208

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem & Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python. Zenodo doi:10.5281/zenodo.1212303

Huang, Baixiang, Canyu Chen & Kai Shu. 2025. Authorship attribution in the era of LLMs: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter* 26(2). 21–43. 10.1145/3715073.3715076

Ishihara, Shunichi. 2021. Score-based likelihood ratios for linguistic text evidence with a bag-of-words model. *Forensic Science International* 327. 110980. 10.1016/j.forsciint.2021.110980

Jawahar, Ganesh, Muhammad Abdul-Mageed & Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. En *28th International Conference on Computational Linguistics (COLING)*, 2296–2309. 10.18653/v1/2020.coling-main.208

Kudryavtseva, Anna & Artyom Kovalevskii. 2025. Comparative statistical analysis of word frequencies in Human-Written and AI-Generated Texts. *Glottometrics* 58. 19–34. 10.53482/2025_58_423

Lai, Uyen, Gurjit S. Randhawa & Paul Sheridan. 2023. Heaps' Law in GPT-Neo Large Language Model Emulated Corpora. En *10th International Workshop on Evaluating Information Access (EVIA)*, 10.20736/0002001352

Langlois-Berthelot, Jean. 2021. When linguistics meets computer science: Stylometry. *Training Language and Culture* 5(2). 51–61. 10.22363/2521-442X-2021-5-2-51-61

Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum & Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences* 28(6). 517–540. 10.1016/j.tics.2024.01.011

Mandelbrot, Benoit. 1953. An information theory of the statistical structure of language. En *Proceedings of the Symposium on Applications of Communication Theory*, 486–502

Mathews, Daniel, Justin Varghese & Libin Chacko Samuel. 2024. Classifying AI-generated summaries and human summaries based on statistical features. En *International Conference on Trends in Quantum Computing and Emerging Business Technologies*, 1–5. 10.1109/TQCEBT59414.2024.10545131

Mitchell, Eric, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning & Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. En *40th International Conference on Machine Learning*, 24950–24962

Moyano-Moreno, Isabel. 2023. La estilometría como técnica de análisis lingüístico-cuantitativo: a propósito del uso terminológico de estilo. *Pragmalingüística* 31. 339–361. 10.25267/Pragmalinguistica.2023.i31.15

Moyano-Moreno, Isabel. 2025. La estilometría computacional desde las humanidades digitales: propuesta metodológica para el análisis estilométrico de novelas anglófonas realistas y modernistas. En *Voces emergentes de la lingüística actual*, 273–292. Editorial Universidad de Sevilla. 10.12795/9788447227570

Nguyen-Son, Hoang-Quoc, Ngoc-Dung T. Tieu, Huy H. Nguyen, Junichi Yamagishi & Isao Echizen. 2017. Identifying computer-generated text using statistical analysis. En *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1504–1511. 10.1109/APSIPA.2017.8282270

Pizarro, Juan. 2019. Using N-grams to detect Bots on Twitter. En *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*,

Queralt, Sheila, Beatriz Esparcia, Marco R. Lessi, Lucía Sánchez-Vecina & Laura Úbeda Cuspinera. 2025. ¿IA, Humano o híbrido?: Fiabilidad de herramientas de detección de IA en textos multiautoría. *INTELETICA* 2(4). 135–149. 10.4114/INTELETICA.vol2iss4pp102-114

Richburg, Aquia, Calvin Bao & Marine Carpuat. 2024. Automatic authorship analysis in human-AI collaborative writing. En *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 1845–1855.

Sarvazyan, Areg Mikael, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi & Paolo Rosso. 2023. Overview of AuTexTification at IberLEF 2023: Detection and attribution of machine-generated text in multiple domains. *Procesamiento del Lenguaje Natural* 71. 275–288. 10.26342/2023-71-21

Savoy, Jacques. 2020. *Machine learning methods for stylometry: Authorship attribution and author profiling*. Cham: Springer. 10.1007/978-3-030-53360-1

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3). 379–423. 10.1002/j.1538-7305.1948.tb01338.x

Sharma, Surbhi & Irfan Mansuri. 2024. Team innovative at SemEval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection. En *18th International Workshop on Semantic Evaluation (SemEval)*, 1172–1176. 10.18653/v1/2024.semeval-1.171

Sumedha & Rohilla. 2021. Empirical laws of natural language processing for neural language generated text. En *Innovative Computing and Communications (ICICCT)*, 184–197. 10.1007/978-3-030-88378-2_15

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* 59(236). 433–460. 10.1093/mind/LIX.236.433

Wales, Katie. 2014. *A dictionary of stylistics*. London: Routledge

Zanotto, Sergio E. & Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. arXiv [cs.CL]. 10.48550/arXiv.2412.03025

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Reading, MA: Addison-Wesley