ISSN: 2386-9658 https://doi.org/10.21071/arf.v37i.18687

# El corpus ROBOT-TALK para el reconocimiento del origen robótico de textos en español

## *(The ROBOT-TALK corpus for recognising the robotic origin of Spanish texts)*

LARA ALONSO SIMÓN
https://orcid.org/0000-0002-7278-9122
laraal04@ucm.es
*Universidad Complutense de Madrid* https://ror.org/02p0gd045

ANA M.ª FERNÁNDEZ-PAMPILLÓN CESTEROS
https://orcid.org/0000-0002-6606-0159
apampi@ucm.es
*Universidad Complutense de Madrid* https://ror.org/02p0gd045

Fecha de recepción: 10 de octubre de 2025
Fecha de aceptación: 26 de diciembre de 2025

**Resumen:** ROBOT-TALK es un corpus monitor comparable de textos humanos en español y su contrapartida escrita por grandes modelos generativos del lenguaje (LLM). Su objetivo es permitir el estudio de posibles rasgos lingüísticos diferenciadores entre textos generados automáticamente y los escritos por las personas. El corpus constituye un recurso lingüístico en español para el reconocimiento de autoría humana vs. "robótica" de textos y diseñado para (1) permitir estudios lingüísticos contrastivos entre LLM y humanos o entre LLM, (2) estudiar la evolución lingüística de los LLM, y (3) servir de soporte en la creación de métodos lingüísticos y herramientas informáticas para la atribución de la autoría humana o automática. Contiene textos de tres géneros diferentes en la lengua escrita (artículos científicos, noticias y reseñas). Cada par de textos, de longitud similar, trata el mismo tema para poder comparar entre dos tipos de escritura y analizar con fiabilidad las características discursivas de los textos. Se recogen muestras de gpt-3, text-davinci-003, babbage-002, curie, gpt-3.5-turbo, gpt-4, bloom, bard, gemini-2.0-flash, gemini-2.5-flash, falcon-180B-chat, Mixtral-8x7B-Instruct-v0.1, claude-3-5-sonnet-20240620, claude-3-7-sonnet-20250219 y DeepSeek-V3. El etiquetado en XML de los textos del corpus permite su consulta con cualquier herramienta de análisis textual que soporte este

estándar de marcado. ROBOT-TALK se ha utilizado con la herramienta SketchEngine para realizar (1) un análisis lingüístico con el fin de encontrar los rasgos más salientes que caracterizan los textos generados por los LLM; (2) un análisis estadístico de rasgos lingüísticos propios de los LLM frente a un posible estilo general humano en español; (3) un análisis lingüístico forense para verificar la fiabilidad en la atribución de autoría; (4) la construcción de clasificadores automáticos binarios y multiclase basados en aprendizaje automático para distinguir textos róboticos y humanos.

**Palabras clave:** Corpus de texto en español. Corpus monitor comparable. Grandes modelos del lenguaje. Atribución de autoría.

**Abstract:** ROBOT-TALK is a comparable corpus of human texts in Spanish and their counterparts written by large language models (LLMs). Its objective is to enable the study of possible linguistic features that differentiate between automatically generated texts and those written by humans. The corpus is a Spanish language resource for recognising human vs. "robotic" authorship of texts and it is designed to (1) enable contrastive linguistic studies between LLMs and humans or between LLMs, (2) study the linguistic evolution of LLMs, and (3) support the creation of linguistic methods and computational tools for attributing human or automatic authorship. It contains texts of three different genres in written language (scientific articles, news articles, and reviews). Each pair of texts, of similar length, deals with the same topic so that the two types of writing can be compared and the discursive characteristics of the texts can be reliably analysed. Samples are collected from gpt-3, text-davinci-003, babbage-002, curie, gpt-3.5-turbo, gpt-4, bloom, bard, gemini-2.0-flash, gemini-2.5-flash, falcon-180B-chat, Mixtral-8x7B -Instruct-v0.1, claude-3-5-sonnet-20240620, claude-3-7-sonnet-20250219 and DeepSeek-V3. The XML tagging of the texts in the corpus allows them to be queried with any text analysis tool that supports this markup standard. ROBOT-TALK has been used with the SketchEngine tool to perform (1) a linguistic analysis to find the most salient features that characterise the texts generated by LLMs; (2) a statistical analysis of linguistic features specific to LLMs compared to a possible general human style in Spanish; (3) a forensic linguistic analysis to verify the reliability of authorship attribution; and (4) the construction of automatic binary and multi-class classifiers based on machine learning to distinguish between robotic and human texts.

**Keywords:** Spanish text corpus. Comparable monitor corpus. Large language models. Authorship attribution.

## 1. ¿Por qué el corpus ROBOT-TALK?

La investigación en la generación del lenguaje natural ha dado lugar a un tipo de modelos generativos de texto (Text Generative Models – TGM) de tamaño muy grande (Large Language Models en inglés) capaces de generar texto con una corrección gramatical, fluidez y coherencia difícil de distinguir de los textos escritos por humanos (Stiff y Johansson 2022; Uchendu, Le y Lee 2023).

Estos actuales grandes modelos generativos del lenguaje (en adelante solo "modelos") están basados en un tipo de red neuronal (*Transformer* con solo decodificador) con miles de millones de conexiones (parámetros) que se ajustan mediante entrenamiento autorregresivo. Así, los modelos resultantes, lo único que "saben hacer" es, dada una secuencia de token (palabras u otro tipo de símbolos), predecir cuál es el siguiente token más probable[1]. Esto los hace especialmente eficaces en aplicaciones como la generación de respuestas conversacionales, generación de resúmenes, documentos o el autocompletado de código entre otras (Jawahar, Abdul-Mageed y Lakshmanan 2020).

El problema surge cuando esta generación de texto de altísima calidad se utiliza con fines maliciosos en situaciones como la manipulación de debates, la difusión de noticias falsas, la generación de reseñas falsas, el plagio o el envío de correos electrónicos de amenaza o extorsión (Stiff y Johansson 2022). Esto plantea desafíos de alto impacto político, económico y social (Pizarro 2019; Pavlyshenko 2022; Cardenuto *et al.* 2023; Crothers, Japkowicz y Viktor 2023).

En este contexto, resulta crítico encontrar métodos fiables para detectar si el autor de un texto es una persona o una máquina (Maloyan, Nutfullin y Ilyushin 2022). Así, conocer la autoría de los textos no sólo ayudará a prevenir y hacer frente a su mal uso, sino que, además, facilitará el cumplimiento del derecho de los ciudadanos a recibir información fiable y transparente (Gobierno de España 2021).

La detección de la autoría humana o automática de un texto es un problema de reconocimiento de autoría no resuelto todavía con suficiente fiabilidad (Wang 2017; Uchendu *et al.* 2020). Nuestro trabajo de investigación se centra en esta cuestión en textos en lengua española[2] y nuestra hipótesis para resolverla es que es posible identificar la autoría humana o por grandes modelos del lenguaje siempre que:

---

[1] En realidad, calculan una distribución de probabilidades sobre todos los token del vocabulario del modelo del lenguaje. El token siguiente es uno de los más probables o el más probable según la estrategia de selección final.

[2] Esta investigación se enmarca en el proyecto ROBOT-TALK. Reconocimiento del origen robótico de textos. Automatización de tareas y conocimiento lingüístico (PID2022-140897OB-I00). Véase https://www.ucm.es/robottalk/

(1) los textos generados por estos modelos tengan unas características lingüísticas propias, algunas de las cuales sean comunes a todos los modelos y

(2) los textos generados por las personas compartan unas características lingüísticas diferentes de las características de los textos generados por los grandes modelos del lenguaje.

Así pues, uno de los objetivos de nuestra investigación es encontrar, si existen, las características lingüísticas que diferencian unos textos de otros. Para ello, es imprescindible disponer de una muestra suficientemente representativa de la escritura de las personas vs. la de los grandes modelos del lenguaje que nos permita realizar los análisis lingüísticos contrastivos para encontrar estas características diferenciadoras. El corpus ROBOT-TALK constituye esta muestra.

## 2. Estado de la cuestión

Centrándonos en la atribución de autoría de un texto, el problema consiste en determinar su verdadero autor de entre un conjunto concreto de posibles autores. Los problemas actuales de la atribución de autoría de textos pueden clasificarse, basándonos en Uchendu, Le y Lee (2023) y Huang, Chen y Shu (2025), del siguiente modo:

1) atribución de autoría de textos humanos. Dado un texto humano, se busca su atribución a un autor concreto de entre varios posibles.

2) detección de textos robóticos. Dado un texto humano o robótico, se persigue determinar si el texto ha sido generado automáticamente por un LLM.

3) atribución de autoría de textos humanos y robóticos. Dado un texto humano o robótico, se trata de precisar el autor del texto, siendo los posibles un LLM dado de un conjunto o un humano.

4) atribución de autoría de textos coescritos por humanos y LLM. Dado un texto humano, robótico o una coautoría entre los dos, se busca concretar el autor del texto, siendo los posibles un humano, un LLM o una coautoría

5) atribución de autoría de textos que han sufrido ocultación de autoría. Dado un texto con ocultación de autoría, se pretende medir el grado de éxito o de evasión de la ocultación.

Para investigar los problemas de atribución de autoría (problema 1) se necesitan corpus con textos dubitados e indubitados. El corpus necesario para resolver el primero de los problemas contendrá exclusivamente textos escritos por humanos. Sin embargo, a fin de resolver el resto de los problemas de autoría mencionados se requiere conjuntos de datos que incluyan texto generado por LLM.

Para dar solución al problema de atribución de autoría de textos coescritos por humanos y LLM (problema 4), el corpus debe incluir textos humanos, textos generados por LLM y también textos de autoría conjunta, como el creado por Dugan *et al.* (2023), que contiene textos cuyas primeras frases proceden de un texto escrito por humanos y las siguientes son una continuación generada por una máquina. En la construcción del corpus de Liyanage, Buscaldi y Nazarenko (2022) se han reemplazado oraciones de resúmenes de artículos científicos humanos por generaciones del modelo Arxiv-NLP[3].

Para llegar a conclusiones en el problema de la atribución de autoría de textos que han sufrido ocultación de autoría (problema 5) se hace necesario que el corpus albergue muestras de textos en los que se haya enmascarado de algún modo el estilo de escritura para ocultar la identidad del autor. Para estos casos encontramos corpus que incluyen textos generados con modelos de parafraseo como el de Rosati (2022), el de Shamardina *et al.* (2022) o el de Sadasivan *et al.* (2024). Antoun *et al.* (2023) han incluido en su corpus ejemplos adversariales escritos por humanos con acceso a ChatGPT y Bing para obtener un estilo similar al de estos generadores. Islam *et al.* (2023) utilizaron ChatGPT para reformular textos humanos del dataset de Kaggle[4], Liu *et al.* (2023) le pidieron que revisara y mejorara un conjunto de artículos académicos, Mindner, Schlippe y Schaaff (2023) lo utilizaron para reescribir textos de Wikipedia y Mitrović, Andreoletti y Ayoub (2023) para reescribir reseñas de restaurantes. Tanto en el trabajo de Mitchell *et al.* (2023) como en el de Lee y Jang (2024) se añadieron perturbaciones o reescrituras con el modelo T5 a los textos robóticos. He *et al.* (2024) añaden en los textos robóticos ataques adversariales de reescritura, introducción de espacios y otras perturbaciones utilizando la librería TextAttack[5].

En el corpus creado por Abdalla *et al.* (2023) se incluye un componente de 4000 textos co-creados por humanos y por reformulaciones del modelo gpt-3.5-turbo. El corpus CHEAT (Yu *et al.* 2025) alberga resúmenes artificiales de artículos de informática creados indicándole a ChatGPT que pula los escritos por humanos y también resúmenes mixtos en los que se mezclan textos escritos por humanos y textos pulidos por ChatGPT. Alhijawi *et al.* (2025) construyen el dataset AIGTxt, que contiene textos generados por ChatGPT mediante reescritura de textos humanos y textos mixtos que combinan el 50 % de textos humanos y el 50 % de textos generados por ChatGPT.

La Tabla 1 resume los corpus encontrados para abordar la atribución de autoría de textos robóticos.

[3] https://huggingface.co/lysandre/arxiv-nlp
[4] https://www.kaggle.com/datasets/rajugc/kaggle-dataset
[5] https://github.com/QData/TextAttack/

| Corpus | Lengua; Dominio | Binario o multiclase | Fuentes humanas | Modelos generadores |
|---|---|---|---|---|
| Guo *et al.* 2023 | inglés y chino; multidominio | binario | inglés (ELI5 dataset, WikiQA dataset, Crawled Wikipedia, Medical Dialog dataset, FiQA dataset); chino (WebTextQA & BaikeQA, Crawled BaiduBaike, NLPCC-DBQA dataset, Medical Dialog dataset, ChineseNlpCorpus, Baidu AI Studio, LegalQA dataset) | ChatGPT |
| Fagni *et al.* 2021 | inglés; tweets | multiclase | tweets de 17 cuentas humanas | Markov Chains, RNN, RNN+Markov, LSTM, GPT-2 |
| Liu *et al.* 2023 | inglés; redacciones de clase de estudiantes de inglés | multiclase | WECCL, TOEFL11 y preparación de exámenes de admisión a posgrado GRE | gpt2-xl, text-babbage-001, text-curie-001, text-davinci-001, text-davinci-002, text-davinci-003, gpt-3.5-turbo |
| Uchendu *et al.* 2021 | inglés; noticias | TuringBenchAA multiclase; TuringBenchTT binario | CNN, Washington Post y Kaggle | GPT-1, GPT-2_small, GPT-2_medium, GPT-2_large, GPT-2_xl, GPT-2_PyTorch, GPT-3, GROVER_base, GROVER_large, GROVER_mega, CTRL, XLM, XLNET_base, XLNET_large, FAIR_wmt19, FAIR_wmt20, |

| | | | | |
|---|---|---|---|---|
| | | | | TRANSFORMER_XL, PPLM_distil, PPLM_gpt2 |
| Doru *et al.* 2025 | alemán; ensayos médicos | binario | ensayos médicos escritos por estudiantes de doctorado | GPT-3.5 |
| Gromov *et al.* 2024 | ruso, alemán, inglés, francés y vietnamita; obras literarias de dominio público | binario | Project Gutenberg y corpus nacionales equivalentes | LSTM, GPT-2, mGPT y YaLM |
| Zaitsu y Jin 2023 | japonés; artículos académicos de psicología | multiclase | The Japanese Journal of Psychology, The Japanese Journal of Criminal Psychology y The Japanese Journal of Social Psychology | GPT-3.5 y GPT-4 |
| Sarvazyan *et al.* 2023 | inglés y español; tweets, reseñas, noticias, legislación y artículos instructivos | multiclase | MultiEURLEX, XSUM, XLSUM, MLSUM, Amazon Reviews, WikiLingua, COAR & COAH, XLM-Tweets, TSATC y TSD | BLOOM-1B7, BLOOM-3B, BLOOM-7B1, GPT-3-babbage, GPT-3-curie y GPT-3-text-davinci-003 |
| Corizzo y Leal-Arenas 2023 | inglés y español; redacciones argumentativas de estudiante de inglés y español | binario | Uppsala Student English Cor- pus (USE) y UC Davis Corpus of Written Spanish, L2 and Heritage Speakers (COWSL2HS) | Chat-GPT |

Tabla 1: Corpus creados con el fin de resolver
la atribución de autoría de textos robóticos

Con el objetivo inicial propuesto en nuestra investigación de descubrir cuáles son los rasgos más salientes de los textos robóticos que puedan ayudar tanto a su detección como a determinar cuál es el LLM generador (problemas 2 y 3 respectivamente), se precisa de un corpus de textos lo más representativo posible del estilo de escritura de las personas (suponiendo que compartimos un estilo común diferente de los LLM) y del estilo (o estilos) de escritura de los LLM. Es en este punto en el que se enmarca este trabajo de diseño y construcción del corpus.

Quedan fuera de nuestra investigación los textos traducidos automáticamente como los contenidos en el corpus SynSciPass (Rosati 2022) o en el corpus RuATD (Shamardina *et al.* 2022). Alshammari, El-Sayed y Elleithy (2024), para la construcción de su corpus en árabe, tradujeron textos generados por ChatGPT en inglés utilizando Google Translate[6].

## 3. Metodología

Para la creación del corpus de textos indubitados ROBOT-TALK utilizamos un modelo evolutivo de tipo espiral frecuentemente usado en Ingeniería del software (Pressman 2010: 39-40). Básicamente el corpus se está creando de forma evolutiva e iterativa siguiendo las fases de:

1) planificación, fase en la que definimos el objetivo del corpus, la planificación temporal, los recursos disponibles y el equipo de trabajo;

2) modelado, en la que hicimos un análisis de requisitos y diseñamos los criterios de diseño, el esquema del corpus —la macroestructura y la microestructura—, así como el procedimiento colaborativo de construcción y validación del corpus;

3) construcción del corpus, fase en la que hemos ido compilando los textos de forma incremental;

4) despliegue del corpus, esta fase se refiere al uso del corpus para los análisis contrastivos o el prototipado de sistemas de identificación automática de autoría;

5) retroalimentación y definición de ajustes, en esta fase recibimos las valoraciones del equipo de investigación que lo utiliza y definimos los ajustes para resolver las limitaciones que detectamos. Estos ajustes conducen a un rediseño del corpus que incorpore los ajustes definidos, volviendo a repetirse iterativamente las siguientes fases de (2) modelado (3) construcción, (4) despliegue y uso, y (5) retroalimentación y ajustes.

Las ventajas que nos ha proporcionado esta metodología de trabajo son fundamentalmente dos: (i) ha facilitado el escalado y la mejora incremental de las

[6] https://translate.google.com/about/?hl=es

sucesivas versiones, lo que es muy adecuado cuando no se conoce a fondo qué tipo de muestras se van a encontrar, y (ii) ha facilitado la adaptación dinámica del esquema y las muestras a la aparición y desaparición en muy cortos periodos de tiempo de los modelos del lenguaje o de sus versiones.

## 4. Objetivos y características del corpus ROBOT-TALK

El objetivo principal del corpus ROBOT-TALK es constituir una muestra suficientemente representativa de la escritura de las personas vs. la de los grandes modelos del lenguaje con el fin de poder estudiar cuáles son las características lingüísticas diferenciadoras entre ambos tipos de texto. Además, un objetivo secundario, pero no menos importante, es servir como corpus para apoyar la construcción de sistemas automáticos o semiautomáticos (clasificadores u otros) de atribución de autoría. Estos objetivos de construcción del corpus nos han guiado en su diseño y son los que determinan las características de ROBOT-TALK.

Atendiendo a la clasificación de tipos de corpus de Rojo (2021), el corpus ROBOT-TALK no es un corpus general en cuanto a que no contiene textos de todo el ámbito hispánico, sino documentos escritos en el español de la variedad de España. En este sentido el corpus es monolingüe por tener textos de una sola lengua, el español. Sin embargo, ROBOT-TALK es un corpus comparable por autores, ya que contiene textos escritos por humanos y sus contrapartidas robóticas generadas de manera automática por diferentes LLM, de modo que las diferencias encontradas entre los textos puedan atribuirse a la diferencia entre los autores. Cada par de textos trata del mismo tema y tiene una longitud similar para poder comparar entre dos tipos de escritura y analizar con fiabilidad las características discursivas de los textos.

Dentro de la distinción entre los corpus de carácter general (o de referencia) y de carácter especializado, el corpus ROBOT-TALK pertenece a esta última categoría, ya que los documentos que lo constituyen están restringidos a un cierto tipo de comunicación (Rojo 2021). En concreto, los textos del corpus pertenecen a tres géneros de diferentes niveles de formalidad en la lengua escrita: artículos científicos de lingüística, noticias y reseñas de cine. La selección de los tres géneros se ha realizado de forma que esté representado el uso del lenguaje conforme diferentes niveles formulaicos y bajo la hipótesis de que habrá patrones lingüísticos que varíen en función del género.

En cuanto a la época en la que se han escrito los textos, el corpus es sincrónico, ya que los textos, de momento, se enmarcan en un margen temporal de no más de veinticinco años (Rojo 2021). Sin embargo, el corpus es un corpus monitor en cuanto que permitirá estudios en diacronía para poder observar el

cambio lingüístico de los modelos generativos (Sinclair 1991). Es un corpus monitor (Sinclair 2005) de la evolución temporal de los LLM, ya que se recogen muestras de los mismos modelos en diferentes momentos y codificamos su fecha producción, así como muestras de diferentes versiones de los modelos también con su fecha exacta de generación. Monitorizar la continua evolución de los LLM permitirá describir el paso del tiempo por los modelos y hacer estudios diacrónicos. Además, el corpus se ha diseñado desde un principio como un corpus abierto de forma que se debe asegurar que:

(1) los textos generados por personas y por robots están públicamente disponibles, no están sometidos a protección de datos, o a aspectos éticos o copyright,

(2) no se parte de un tamaño preestablecido para que puedan ir incorporándose nuevos textos en el tiempo a medida que otros nuevos modelos de lenguaje generativo vayan estando disponibles.

En la Tabla 2 recogemos los modelos del lenguaje que hemos utilizado para la generación de los textos automáticos, junto con la fecha en que empezamos a utilizarlos y la fecha fin en la que dejamos de tener acceso a ellos. Los modelos se han seleccionado aplicando dos criterios: que generen textos de muy alta calidad y sean accesibles mediante API o una interfaz. El corpus ROBOT-TALK actualmente contiene textos de gpt-3, text-davinci-003, babbage-002, curie, gpt-3.5-turbo y gpt-4 de OpenAI[7]; bloom de BigScience[8]; bard, gemini-2.0-flash y gemini-2.5-flash de Google[9]; falcon-180B-chat de Technology Innovation Institute[10]; Mixtral-8x7B-Instruct-v0.1 de Mistral AI[11]; claude-3-5-sonnet-20240620 y claude-3-7-sonnet-20250219 de Claude[12]; y DeepSeek-V3 de Deep Seek[13].

| CÓDIGO DEL LLM | FUENTE | VERSIÓN |
|---|---|---|
| gpt3 | OpenAI | gpt-3 |
| txdv | OpenAI | text-davinci-003 |
| babb | OpenAI | babbage-002 |
| blom | BigScience | bloom |

[7] https://openai.com
[8] https://bigscience.huggingface.co
[9] https://gemini.google.com
[10] https://www.tii.ae
[11] https://mistral.ai
[12] https://www.anthropic.com
[13] https://www.deepseek.com

| curi | OpenAI | curie |
|---|---|---|
| g35t | OpenAI | gpt-3.5-turbo |
| gpt4 | OpenAI | gpt-4 |
| bard | Google | bard |
| falc | TII | falcon-180B-chat |
| mxit | MistralAI | Mixtral-8x7B-Instruct-v0.1 |
| clau | Anthropic | claude-3-5-sonnet-20240620 |
| bard | Google | gemini-2.0-flash |
| deep | DeepSeek | DeepSeek-V3 |
| clau | Anthropic | claude-3-7-sonnet-20250219 |
| bard | Google | gemini-2.5-flash |

Tabla 2: LLM del corpus y fechas de recogida

En lo referente al medio o soporte material de los documentos, el corpus contiene únicamente documentos escritos en formato electrónico (Sinclair 2003). En cuanto a la codificación de los textos, en cada uno de ellos se ha añadido una cabecera con los metadatos de información extratextual congruente con la organización del corpus (Rojo 2021) que nos permitirán recuperar la definición del lugar que ostenta cada texto en el corpus y también distinguir de forma clara el texto del documento de la información asociada. Los documentos del corpus se han almacenado en texto plano y se han codificado en XML (Extended Mark-up Language), lo que posibilita su consulta con cualquier herramienta de análisis textual (ej. SketchEngine[14]) que soporte este estándar de marcado. En la siguiente sección se detalla el esquema de codificación del corpus.

## 5. El esquema del corpus

### *5.1. La macroestructura*

El corpus ROBOT-TALK se organiza, a nivel de macroestructura, en dos niveles jerárquicos que determinan dos componentes contrastivos: por autor y por género. La correspondencia entre los textos comparables se codifica en el nombre de los documentos XML de forma que sea fácilmente interpretable por los humanos. Este nombre, que coincide con el identificador unívoco de cada muestra, indica el género, el autor y un número entero único para cada texto. La Figura 1 muestra la macroestructura del corpus ROBOT-TALK.

[14] https://www.sketchengine.eu/

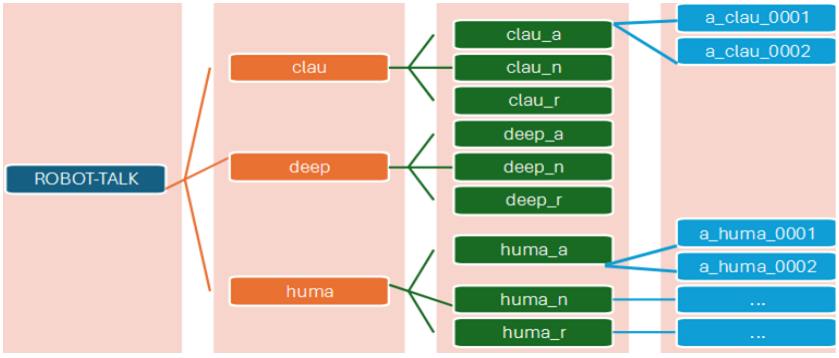


Figura 1: Macroestructura del corpus ROBOT-TALK

*5.2. Microestructura*

En cuanto a la microestructura de ROBOT-TALK, cada muestra del corpus se almacena en un archivo de texto en formato XML con los metadatos y la estructura del contenido textual dependiente de cada género. Cada documento XML que se crea de un texto se valida con el esquema XSD antes de su almacenamiento.

5.2.1. Metadatos del corpus ROBOT-TALK

Los metadatos tienen el objetivo de permitir el filtrado de la información en las búsquedas automáticas para favorecer la explotación del corpus y la comparabilidad entre componentes contrastivos. La Figura 2 muestra un ejemplo de metadatos.

```
id="a_huma_0011"
id_refs="a_bard_0011 a_clau_0011 a_deep_0011 a_falc_0011 a_g35t_0011
a_gpt4_0011 a_mxit_0011"
tipo_version="humano"
contribuidor="Lara"
fuente="Revista electrónica de lingüística aplicada"
autor="Ruth María Lavale Ortiz"
fecha="20130925"
url="https://dialnet.unirioja.es/servlet/articulo?codigo=4648343"
dominio="artículos"
ambito="lingüística"
titulo="La formación de verbos denominales por derivación: morfología
y semántica en la clase de ELE"
palabras_clave="morfología, semántica, verbos denominales, ELE"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="corpus_comparable.xsd">
```

Figura 2: Metadatos de los textos del corpus ROBOT-TALK

El atributo *id* es el identificador unívoco del documento y coincide con el nombre del archivo. Contiene los datos del género textual y del tipo de autor del texto.

El *id_ref* en los textos automáticos tiene como valor el *id* del texto humano a partir del que se ha generado el texto automático. Paralelamente el *id_refs* en los metadatos de los textos humanos contiene todos los *id* de los textos automáticos que se han producido a partir de ese texto humano. Los atributos *idref* e *idrefs* permiten codificar la comparabilidad entre los textos de diferentes autores.

Otros atributos recogen la información relativa al origen del texto (tipo versión, fuente, autor, fecha y url) y al género textual (dominio, ámbito, subámbito). Asimismo, la monitorización de la evolución temporal de los LLM la codificamos con los atributos *tipo_version* y la *fecha* de producción.

Finalmente, en los metadatos también se incluyen los siguientes datos: título y las palabras clave en los artículos científicos; el título en las noticias; y el título de la obra que se reseña, el título de la reseña y la puntuación otorgada en las reseñas. Además, se refleja la identidad del contribuidor que se ha encargado de procesar el documento completo, las recomendaciones W3C bajo las que se ha formado el documento XML y el esquema XSD con el que se valida el documento.

5.2.2 Estructura del contenido textual

En cuanto a la estructura del contenido textual de cada género textual, en los artículos las partes de texto plano que se almacenan son el resumen, la introducción y las conclusiones del artículo para asegurar la comparabilidad entre

los diferentes artículos; ya que estas son tres secciones que tienen todos los artículos (Figura 3). En las noticias y reseñas únicamente se almacena el texto plano en la etiqueta *cuerpo*.

```
<resumen>En este artículo ofrecemos una reflexión sobre la relevancia
de la enseñanza de las reglas de formación de palabras en la clase de
español como lengua extranjera. La morfología es la gran olvidada en
las clases de español, a pesar de [...]</resumen>
<introduccion>Han sido muchos los profesores de español e
investigadores en morfología que han puesto de manifiesto que la
morfología —junto a la fonética— es una de las grandes olvidadas en las
clases de español como lengua [...]</introduccion>
<conclusiones>La libertad del hablante a la hora de crear verbos
formados sobre sustantivos es amplia, pero los patrones morfológicos
son limitados —derivados y parasintéticos, aunque aquí nos hemos
centrado únicamente en los más [...]</conclusiones>
```

Figura 3: Etiquetado de la estructura del contenido de un artículo científico

## 6. Procedimiento de construcción del corpus ROBOT-TALK

El procedimiento de creación del corpus ROBOT-TALK consta de cuatro fases: (1) la búsqueda del texto humano, (2) el almacenamiento del texto humano, (3) la generación del texto robótico comparable y (4) el almacenamiento del texto robótico. La Figura 4 refleja el diagrama del proceso.

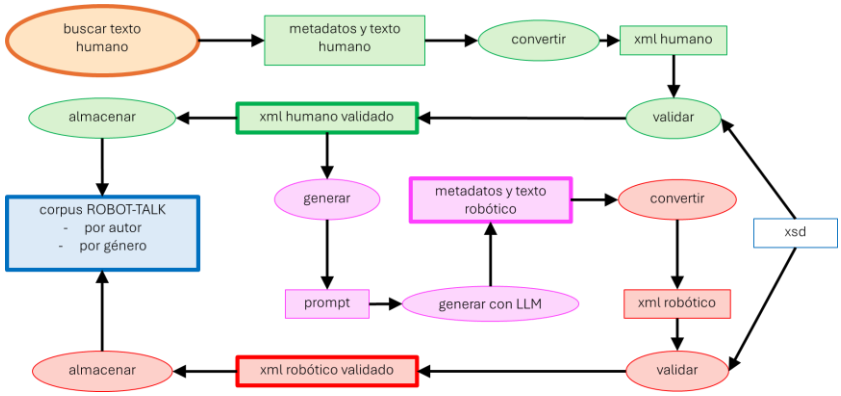


Figura 4: Diagrama del proceso de construcción

### *6.1. Búsqueda del texto humano*

En la primera fase se localizan los textos humanos a partir de los cuales se generarán todas las contrapartidas automáticas. Esta búsqueda se hace de forma manual asegurando que la autoría del texto es humana. Para este propósito, los artículos se buscan en revistas científicas de lingüística españolas en abierto de

reconocido prestigio que tengan revisión por pares. Las revistas se seleccionan por la facilidad de la extracción del texto. Las noticias se seleccionan de las webs de noticias RTVE[15] y EFE[16]; estas webs son fuentes de reconocido prestigio, no necesitan suscripción y el texto es fácilmente extraíble. Las reseñas de cine se recogen de la web de Filmaffinity[17], una página en abierto donde los usuarios escriben reseñas de películas y series otorgándoles además una puntuación entre 1 y 10. Se seleccionan reseñas de usuarios de España.

*6.2. Almacenamiento del texto humano en formato XML*

En la segunda fase se crea, valida y almacena el documento XML con el contenido del texto humano buscado (Figura 5). Esta fase es automática y se puede hacer para cada documento o bien para varios documentos a la vez a partir de un archivo csv con el texto y los valores de los atributos de los metadatos.

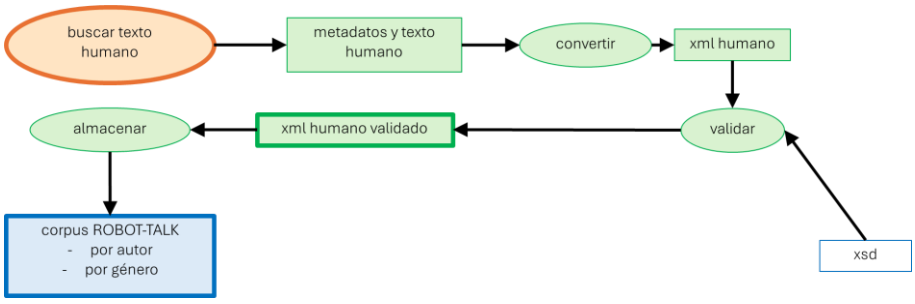


Figura 5: Paso 2: almacenamiento de XML humano

*6.3. Generación del texto robótico comparable*

En la tercera fase generamos los textos automáticos comparables mediante instrucciones o *prompts*. Este es un paso semiautomático, ya que creamos el *prompt* de forma automática, pero alimentamos el LLM de forma manual (Figura 6).

[15] https://www.rtve.es
[16] https://efe.com
[17] https://www.filmaffinity.com/es/main.html

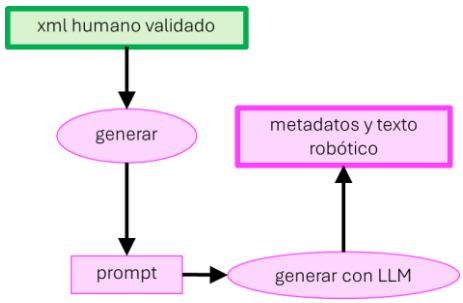


Figura 6: Paso3: creación del texto robótico comparable

El *prompt* se genera de forma automática, a partir del XML humano validado. Recoge las instrucciones de formato, fija el rol que el LLM debe tomar, proporciona el contexto, el tema a tratar y la extensión en número de palabras que debe tener el texto. Cada texto automático se genera en una nueva sesión de la interfaz del modelo tras borrar el contexto anterior para asegurar que no haya interferencias con instrucciones anteriores.

En la Figura 7 se muestra la plantilla de generación del *prompt* para el resumen de los artículos científicos. Para asegurar la coherencia entre las secciones utilizamos: (i) para el resumen, el título y las palabras clave del artículo humano; (ii) para la introducción, el título y las conclusiones, y (iii) para las conclusiones, el título y el resumen.

En lo que respecta al género de las noticias, la información textual proporcionada es la sección del periódico a la que pertenece la noticia (cultura, deportes…) y el título. Respecto del género de las reseñas, se informa de si la obra reseñada es una película o una serie y se proporciona el título de la obra, el título de la reseña origen y la polaridad de la opinión en forma de puntuación del 1 al 10.

```
f'''
  Sigue las siguientes instrucciones:
  - Eres un lingüista español y escribes para una revista científica de
  investigación en español del ámbito de la lingüística.
  - Tu tarea es escribir el apartado del resumen de un artículo científico con los
  siguientes datos:
  - Título del artículo: "{articulo["titulo"]}".
  - Palabras clave: "{articulo["palabras_clave"]}".
  - El resumen debe tener una extensión de {articulo["extension_resumen"]} palabras.
  - No escribas nada en negrita.
  - No escribas ningún título.
  - No escribas las palabras clave.
  - No devuelvas el número de palabras que has escrito.
  - No menciones el título del artículo que te he dado.
  - Escribe solo el apartado del resumen.
  - No escribas nada dirigiéndote a mí.
  '''
```

Figura 7: Prompt para la sección de resumen de los artículos

*6.4. Almacenamiento del texto robótico en formato XML*

En la cuarta y última fase, se crea, valida y almacena de forma automática el texto de los LLM en formato XML con los metadatos y la estructura correspondientes de forma análoga a como lo hacemos para los textos humanos (Figura 8).

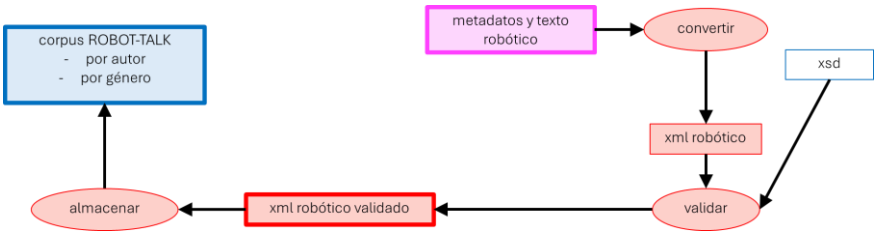


Figura 8: Paso 4: almacenamiento del XML robótico

## 7. El entorno de trabajo

El entorno de trabajo está formado por dos aplicaciones: (i) Sketch Engine, una herramienta de análisis textual a la que se pude acceder bien mediante su interfaz web o bien mediante su API; (ii) Google Colaboratory[18], para la codificación y ejecución de los algoritmos que automatizan el proceso. El lenguaje de programación es Python, en su versión 3.11.13.

## 8. Resultados: el corpus

El corpus, en su estado actual, se ha creado entre diciembre del 2022 y mayo del 2025. Se puede consultar una muestra en:

[18] https://colab.research.google.com/

https://www.ucm.es/robottalk/corpus-robot-talk.

La Tabla 3 muestra la distribución del corpus por tipo de autor y género textual considerando únicamente los siete LLM que ya hemos usado en diferentes análisis del proyecto. El corpus sigue en desarrollo, incorporando nuevos LLM y nuevos textos. Está prevista la publicación en abierto de una parte del corpus, al finalizar el proyecto, en un repositorio institucional.

| **corpus** | **huma** | **bard** | **clau** | **deep** | **falc** | **g35t** | **gpt4** | **mxit** | **total por género** |
|---|---|---|---|---|---|---|---|---|---|
| artículos | 152 | 152 | 152 | 152 | 12 | 90 | 144 | 90 | 944 |
| noticias | 182 | 182 | 182 | 182 | 33 | 111 | 171 | 111 | 1154 |
| reseñas | 171 | 171 | 171 | 171 | 16 | 95 | 160 | 95 | 1050 |
| Total por autor | 505 | 505 | 505 | 505 | 61 | 296 | 475 | 296 | 3148 |

Tabla 3: Distribución por género y autor del corpus ROBOT-TALK

*8.1. Aplicaciones del corpus*

El ROBOT-TALK se ha utilizado para realizar:

1) un análisis lingüístico cualitativo en el que se han encontrado rasgos lingüísticos propios de los textos automáticos y diferentes de los textos humano en todos los niveles lingüísticos: morfológico, léxico, sintáctico, semántico y discursivo. Se ha comprobado también que estos rasgos dependen de los géneros textuales (Alonso Simón, Escandell-Vidal y Fernández Trinidad 2025).

2) Un análisis estadístico de una parte (concretamente 41) de los rasgos lingüísticos encontrados en el análisis lingüístico cualitativo anterior para los que se ha verificado estadísticamente su capacidad discriminatoria de la autoría (Alonso Simón *et al.* 2025; Alonso Simón, Jiménez-Bravo Bonilla y Márquez Cruz 2025).

3) Un análisis lingüístico forense para verificar la fiabilidad en la atribución de autoría utilizando una parte del corpus como textos indubitados y la otra como textos dubitados de evaluación (Crespo Miguel, Moyano Moreno y Bernal Ortiz 2025).

4) La construcción de clasificadores automáticos binarios y multiclase basados en aprendizaje automático para distinguir textos automáticos y humanos (Khalil y Fernández-Pampillón Cesteros 2025).

## Conclusiones

El corpus ROBOT-TALK constituye, hasta donde sabemos, la primera muestra sobre cómo escriben los LLM vs. las personas en español. Es un corpus comparable y monitor que está diseñado para (i) permitir estudios lingüísticos contrastivos entre LLM y humanos o LLM entre sí, (ii) incorporar los textos que generan las nuevas versiones o los nuevos LLM que, de forma rápida y continuada, están sucediendo, para así tener la posibilidad de estudiar la evolución lingüística de los LLM, y (iii) facilitar la creación de métodos lingüísticos y de herramientas informáticas para la atribución de la autoría humana o automática.

El corpus, sin embargo, necesitaría ampliarse en dos sentidos: (1) la incorporación de nuevos dominios de conocimiento en los artículos científicos y de nuevas fuentes de reseñas, con el fin de que se pueda verificar que los rasgos lingüísticos encontrados en cada uno de estos géneros son verdaderamente dependientes del género, y (2) la incorporación de un mayor número de textos para la utilización del corpus para crear y evaluar herramientas automáticas de atribución de autoría. En este sentido, resulta crítico realizar la automatización de todos los pasos del proceso de construcción de forma que se puedan incorporar, con el mínimo coste y en el menor tiempo posible, los textos indubitados de nuevos LLM que puedan ser objetivo de la atribución. En estas dos direcciones, la automatización y ampliación del corpus, se dirige nuestro trabajo futuro.

## Agradecimientos

Este trabajo 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 Gonzalo Ángel-Cruz Burguillo, Iván Arias Rodríguez, José Antonio Gonzalo Gimeno e Icíar Nieva Soto su contribución en la construcción del corpus ROBOT-TALK.

## Referencias bibliográficas

ABDALLA, Mohamed Hesham Ibrahim *et al.*, "A Benchmark Dataset to Distinguish Human-Written and Machine-Generated Scientific Papers". En: *Information*, 14, 522, 2023. Disponible en línea en: https://doi.org/10.3390/info14100522

ALHIJAWI, Bushra *et al.*, "Deep Learning Detection Method for Large Language Models-Generated Scientific Content". En: *Neural Computing and Applications*, 37, 2025, pp. 91-104. Disponible en línea en:

https://doi.org/10.1007/s00521-024-10538-y

ALONSO SIMÓN, Lara; ESCANDELL-VIDAL, M. Victoria; FERNÁNDEZ TRINIDAD, Marianela, "La identificación de textos "robóticos" en español mediante análisis lingüístico". En: Albujer Lax, Miguel Ángel; Soriano Moreno, Claudia (eds.), *XVI Congreso Internacional de Lingüística General (2025). Libro de resúmenes*. Alicante: Universidad de Alicante, 2025, pp. 142-143. Disponible en línea en: https://cilg2025.ua.es/uploads/site/files/Libro_Resumenes_CILG25.pdf

ALONSO SIMÓN, Lara *et al.*, "¿Tienen GPT-3.5 y GPT-4 un estilo de escritura diferente del estilo humano? Un estudio exploratorio para el español". En: *Revista Electrónica de Lingüística Aplicada*, 23, 1, 2025, pp. 34-54. Disponible en línea en: https://doi.org/10.58859/rael.v23i1.666

ALONSO SIMÓN, Lara; JIMÉNEZ-BRAVO BONILLA, Miguel; MÁRQUEZ CRUZ, Manuel, "Reconocimiento de autoría de textos generados por modelos GPT y textos humanos: análisis cuantitativo del estilo con el corpus ROBOT-TALK". En: Albujer Lax, Miguel Ángel; Soriano Moreno, Claudia (eds.), *XVI Congreso Internacional de Lingüística General (2025). Libro de resúmenes.* Alicante: Universidad de Alicante, 2025, pp. 144-145. Disponible en línea en: https://cilg2025.ua.es/uploads/site/files/Libro_Resumenes_CILG25.pdf

ALSHAMMARI, Hamed; EL-SAYED, Ahmed; ELLEITHY, Khaled, "AI-Generated Text Detector for Arabic Language Using Encoder-Based Transformer Architecture". En: *Big Data and Cognitive Computing*, 8, 32, 2024. Disponible en línea en: https://doi.org/10.3390/bdcc8030032

ANTOUN, Wissam *et al.*, "Towards a Robust Detection of Language Model-Generated Text: Is ChatGPT that Easy to Detect?". En: *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, 1, 2023, pp. 14-27. Disponible en línea en: https://aclanthology.org/2023.jeptalnrecital-long.2/

CARDENUTO, João Phillipe *et al.*, "The Age of Synthetic Realities: Challenges and Opportunities". En: *arXiv:2306.11503*, 2023. Disponible en línea en: https://doi.org/10.48550/arXiv.2306.11503

CORIZZO, Roberto; LEAL-ARENAS, Sebastian, "A Deep Fusion Model for Human vs. Machine-Generated Essay Classification". En: *International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1-10. Disponible en línea en: https://doi.org/10.1109/IJCNN54540.2023.10191322

CRESPO MIGUEL, Mario; MOYANO MORENO, Isabel; BERNAL ORTIZ, Francisco, "Análisis de autoría en textos robóticos: enfoque lingüístico-forense mediante técnicas cualitativas y cuantitativas". En: Albujer Lax, Miguel Ángel; Soriano Moreno, Claudia (eds.), *XVI Congreso Internacional de*

*Lingüística General (2025). Libro de resúmenes*. Alicante: Universidad de Alicante, 2025, pp. 148-149. Disponible en línea en: https://cilg2025.ua.es/uploads/site/files/Libro_Resumenes_CILG25.pdf

CROTHERS, Evan N.; JAPKOWICZ, Nathalie; VIKTOR, Herna L., "Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods". En: *IEEE Access*, 11, 2023, pp. 70977-71002. Disponible en línea en: https://doi.org/10.1109/ACCESS.2023.3294090

DORU, Berin *et al.*, "Detecting Artificial Intelligence–Generated Versus Human-Written Medical Student Essays: Semirandomized Controlled Study". En: *JMIR Medical Education*, 11, 2025. Disponible en línea en: https://doi.org/10.2196/62779

DUGAN, Liam *et al.*, "Real or Fake Text?: Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text". En: *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 11, 2023, pp. 12763-12771. Disponible en línea en: https://doi.org/10.1609/aaai.v37i11.26501

FAGNI, Tiziano *et al.*, "TweepFake: About detecting deepfake tweets". En: *PLoS ONE*, 16, 5, 2021. Disponible en línea en: https://doi.org/10.1371/journal.pone.0251415

GOBIERNO DE ESPAÑA, *Carta de Derechos Digitales*, 2021. Disponible en línea en: https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf

GROMOV, Vasilii A. *et al.*, "Spot the bot: the inverse problems of NLP". En: *PeerJ Computer Science*, 10, 2024. Disponible en línea en: https://doi.org/10.7717/peerj-cs.2550

GUO, Biyang *et al.*, "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection". En: *arXiv:2301.07597*, 2023. Disponible en línea en: https://doi.org/10.48550/arXiv.2301.07597

HE, Xinlei *et al.*, MGTBench: "Benchmarking Machine-Generated Text Detection". En: *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 2251-2265. Disponible en línea en: https://doi.org/10.1145/3658644.3670344

HUANG, Baixiang; CHEN, Canyu; SHU, Kai, "Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges". En: *SIGKDD Explor. Newsl*, 26, 2, 2025, pp. 21-43. Disponible en línea en: https://doi.org/10.1145/3715073.3715076

ISLAM, Niful *et al.*, "Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning". En: *arXiv:2306.01761v1*, 2023. Disponible en línea en: https://doi.org/10.48550/arXiv.2306.01761

JAWAHAR, Ganesh; ABDUL-MAGEED, Muhammad; LAKSHMANAN, Laks V. S., "Automatic Detection of Machine Generated Text: A Critical Survey". En: *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 2296-2309. Disponible en línea en: https://doi.org/10.18653/v1/2020.coling-main.208

KHALIL, Doaa Samy; FERNÁNDEZ-PAMPILLÓN CESTEROS, Ana María, "Clasificación automática de la autoría robótica o humana de textos en español". En: Albujer Lax, Miguel Ángel; Soriano Moreno, Claudia (eds.), *XVI Congreso Internacional de Lingüística General (2025). Libro de resúmenes*. Alicante: Universidad de Alicante, 2025, pp. 146-147. Disponible en línea en: https://cilg2025.ua.es/uploads/site/files/Libro_Resumenes_CILG25.pdf

LEE, Dong Hee; JANG, Beakcheol, "Enhancing Machine-Generated Text Detection: Adversarial Fine-Tuning of Pre-Trained Language Models". En: *IEEE Access*, 12, 2024, pp. 65333-65340. Disponible en línea en: https://doi.org/10.1109/ACCESS.2024.3396820

LIU, Yikang *et al.*, "ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models". En: *arXiv:2304.07666v2*, 2023. Disponible en línea en: https://doi.org/10.48550/arXiv.2304.07666

LIU, Zeyan *et al.*, "Check Me If You Can: Detecting ChatGPT-Generated Academic Writing using CheckGPT". En: *arXiv:2306.05524*, 2023. Disponible en línea en: https://doi.org/10.48550/arXiv.2306.05524

LIYANAGE, Vijini; BUSCALDI, Davide; NAZARENKO, Adeline, "A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications". En: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 2022, pp. 4692-4700. Disponible en línea en: https://aclanthology.org/2022.lrec-1.501/

MALOYAN, Narek; NUTFULLIN, Bulat; ILYUSHIN, Eugene, "DIALOG-22 RuATD Generated Text Detection". En: *Computational Linguistics and Intellectual Technologies*, 21, 2022, pp. 396-401. Disponible en línea en: https://scispace.com/pdf/dialog-22-ruatd-generated-text-detection-1dw3v7x9.pdf

MINDNER, Lorenz; SCHLIPPE, Tim; SCHAAFF, Kristina, "Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT". En: Schlippe, Tim; Cheng, Eric C. K.; Wang, Tianchong (eds.), *Artificial Intelligence in Education Technologies: New Development and Innovative Practices. AIET 2023. Lecture Notes on Data Engineering and Communications Technologies*. Singapur: Springer Nature Singapore, 190, 2023, pp. 152-170. Disponible en línea en: https://doi.org/10.1007/978-981-99-7947-9_12

MITCHELL, Eric *et al.*, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature". En: *Proceedings of Machine Learning Research*, 202, 2023, pp. 24950-24962. Disponible en línea en: https://proceedings.mlr.press/v202/mitchell23a/mitchell23a.pdf

MITROVIĆ, Sandra; ANDREOLETTI, Davide; AYOUB, Omran, "ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text". En: *arXiv:2301.13852v1*, 2023. Disponible en línea en: https://doi.org/10.48550/arXiv.2301.13852

PAVLYSHENKO, Bohdan M., "Methods of Informational Trends Analytics and Fake News Detection on Twitter". En: *arXiv:2204.04891v1*, 2022. Disponible en línea en: https://doi.org/10.48550/arXiv.2204.04891

PIZARRO, Juan, "Using N-grams to detect Bots on Twitter Notebook for PAN at CLEF". En: *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, 2019. Disponible en línea en: https://ceur-ws.org/Vol-2380/paper_183.pdf

PRESSMAN, Roger S., *Ingeniería del software. Un enfoque práctico*. México D. F.: McGraw-Hill, 2010.

ROJO, Guillermo, *Introducción a la lingüística de corpus en español*. Londres: Routledge, Taylor & Francis Group, 2021. Disponible en línea en: https://doi.org/10.4324/9781003119760

ROSATI, Domenic, "SynSciPass: detecting appropriate uses of scientific text generation". En: *Proceedings of the Third Workshop on Scholarly Document Processing*, 2022, pp. 214-222. Disponible en línea en: https://aclanthology.org/2022.sdp-1.27/

SADASIVAN, Vinu Sankar *et al.*, "Can AI-Generated Text be Reliably Detected?" En: *arXiv:2303.11156v4*, 2024. Disponible en línea en: https://doi.org/10.48550/arXiv.2303.11156

SARVAZYAN, Areg Mikael *et al.*, "Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains". En: *Procesamiento del Lenguaje Natural*, 71 2023, pp. 275-288. Disponible en línea en: https://doi.org/10.26342/2023-71-21

SHAMARDINA, Tatiana *et al.*, "Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian". En: *Computational Linguistics and Intellectual Technologies*, 21, 2022, pp. 497-511. Disponible en línea en: https://scispace.com/pdf/findings-of-the-the-ruatd-shared-task-2022-on-artificial-10uxwywq.pdf

SINCLAIR, John, *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

——, "Corpora for lexicography". En: van Sterkenberg, Piet (ed.), *A practical guide to lexicography*. Amsterdam: John Benjamins Publishing Company, 2003, pp. 167-178.

——, "Corpus and text-basic principles". En: Wynne, Martin (ed.), *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 2005, pp. 1-16

STIFF, Harald; JOHANSSON, Fredrik, "Detecting computer-generated disinformation". En: *International Journal of Data Science and Analytics*, 13, 2022, pp. 363-383. Disponible en línea en: https://doi.org/10.1007/s41060-021-00299-5

UCHENDU, Adaku; LE, Thai; LEE, "Dongwon, Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective". En: *ACM SIGKDD Explorations Newsletter*, 25, 1, 2023, pp. 1-18. Disponible en línea en: https://doi.org/10.1145/3606274.3606276

UCHENDU, Adaku *et al.*, "Authorship Attribution for Neural Text Generation". En: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 2020, pp. 8384-8395. Disponible en línea en: https://doi.org/10.18653/v1/2020.emnlp-main.673

UCHENDU, Adaku *et al.*, "TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation". En: *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, 2021, pp. 2001-2016 Disponible en línea en: https://doi.org/10.18653/v1/2021.findings-emnlp.172

WANG, William Yang, ""Liar, liar pants on fire": A new benchmark dataset for fake news detection". En: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2, 2017, pp. 422-426. Disponible en línea en: https://doi.org/10.18653/v1/P17-2067

YU, Peipeng *et al.*, "CHEAT: A Large-scale Dataset for Detecting ChatGPT-writtEn AbsTracts". En: *IEEE Transactions on Big Data*, 11, 3, 2025, pp. 898-906. Disponible en línea en: https://doi.org/10.1109/TBDATA.2025.3536929

ZAITSU, Wataru; JIN, Mingzhe, "Distinguishing ChatGPT (-3.5, -4)-generated and human-written papers through Japanese stylometric analysis". En: *PLoS ONE*, 18, 8, 2023. Disponible en línea en: https://doi.org/10.1371/journal.pone.0288453