CUADERNOS DE DOCENCIA UNIVERSITARIA 54

La intersección entre la IA y el análisis lingüístico

Experiencias de evaluación de la IA en la descripción y el análisis lingüísticos del inglés y el catalán

Natalia Judith Laso Martín
Elisabet Comelles Pujadas
(coordinación)

Octaedro Editorial

Institut de Desenvolupament Professional
UNIVERSITAT DE BARCELONA

Cuadernos de docencia universitaria 54

Título: *La intersección entre la IA y el análisis lingüístico. Experiencias de evaluación de la IA en la descripción y el análisis lingüísticos del inglés y el catalán*

Primera edición: marzo de 2026

Recepción del original: 03/06/25
Aceptación: 22/08/25

Ediciones OCTAEDRO
Bailén, 5, pral. - 08010 Barcelona
Tel.: 93 246 40 02
www.octaedro.com - octaedro@octaedro.com

Universitat de Barcelona
Institut de Desenvolupament Professional
Passeig de la Vall d'Hebron, 171, 08035 Barcelona
Tel.: 934035175
www.ub.edu/idp/web/ - idp.ice@ub.edu

ISBN: 978-84-1079-313-2

Diseño y producción: Servicios Gráficos Octaedro

AUTORÍA

Natalia Judith Laso Martín y Elisabet Comelles Pujadas (coords.), Júlia Malet Raich, Ionela Valentina Vasilovici y Èlia Vilà Colomé

AGRADECIMIENTOS

Queremos expresar nuestro agradecimiento a la Dra. M. Àngels Massip por su valiosa codirección del trabajo *Efectividad de ChatGPT en la revisión y corrección de textos*. Así mismo, agradecemos al proyecto *Making the CEFR/CV more user-friendly: fine-tuning descriptors with Learner Corpus Research (LCR) results* (PID2020-117041GA-I00) y al Servei de Tecnologia Lingüística (STeL) de la Universitat de Barcelona el habernos permitido utilizar el FineDesc Learner Corpus y el corpus GRERLI, respectivamente. Ambos recursos han sido herramientas fundamentales para el desarrollo de este estudio.

ÍNDICE

PRÓLOGO

Últimamente, hemos observado un crecimiento exponencial en el uso de la inteligencia artificial (IA) en diversas áreas del conocimiento, lo que también es evidente en la lingüística y la enseñanza de lenguas. Una de las consecuencias más notables de esta expansión tecnológica es el progresivo aumento de propuestas de trabajos de fin de grado (TFG) que abordan la evaluación del lenguaje generado por IA. Esta tendencia nos ha llevado al grupo de innovación docente consolidado GraLiAng a considerar la incorporación de esta área como una línea de investigación de los TFG que supervisamos en los estudios de Filología.

El presente volumen recoge tres experiencias desarrolladas en el marco de estos TFG en los estudios de Filología, donde se plantean dos aspectos cruciales: por un lado, la potencialidad de diversas herramientas de IA para brindar retroalimentación correctiva de la producción escrita de aprendices de lenguas, y, por otro, la capacidad de diversos bots de conversación para detectar rasgos lingüísticos tan complejos como la ironía.

Estos dos ejes se han seleccionado porque han sido los más tratados por los estudiantes en los TFG recientemente presentados, dentro de una línea de investigación más amplia impulsada por las coordinadoras del volumen, centrada en la evaluación del lenguaje generado por la IA (*Evaluation of AI-generated language*). Aunque, *a priori*, la detección de la ironía y la retroalimentación correctiva de textos académicos pueden parecer ámbitos distantes, ambos responden a un mismo interés: analizar hasta qué punto las herramientas de IA pueden comprender, reproducir o evaluar fenómenos lingüísticos complejos y, a partir de ahí, convertirse en un apoyo real en la descripción y el análisis lingüísticos.

Este interés reciente por la evaluación del uso de la IA aplicada a la descripción y análisis lingüísticos se debe, en parte, al hecho de que estas tecnologías han comenzado a ocupar un lugar primordial como herramientas de apoyo en infinidad de tareas académicas de los alumnos de Filología, tales como la redacción de textos académicos, la corrección de errores ortográficos, gramaticales, la mejora del estilo de escritura académico, así como la generación de contenidos. En este sentido, el de-

sarrollo de herramientas de IA generativas que proporcionan resultados que a menudo podrían pasar por producciones humanas plantea un conjunto de preguntas fundamentales en el campo de la enseñanza y el aprendizaje de lenguas: ¿es posible que una herramienta de IA logre el mismo nivel de comprensión y habilidad para identificar y analizar elementos lingüísticos, como el humor o la ironía? ¿Puede una herramienta de IA ofrecer retroacción correctiva de un texto adaptada a distintos niveles de competencia lingüística y a diferentes tipologías textuales?

Los docentes que hemos impulsado esta línea de investigación no solo vemos en la evaluación del uso de la IA aplicada a la descripción y el análisis lingüísticos una oportunidad académica, sino también una necesidad. El avance tecnológico no es solo una herramienta potencial para ayudar en el aprendizaje, sino que también genera nuevos retos que se han de afrontar con rigor académico. En particular, la IA aplicada a la descripción y el análisis lingüísticos supone una herramienta poderosa para explorar fenómenos lingüísticos de forma sistemática y con un alto grado de precisión (Ettinger et al., 2023; Uchida, 2024; Llu et al., 2024). Sin embargo, todavía hay debates abiertos sobre qué aspectos de la producción lingüística pueden ser eficazmente reproducidos o identificados por una IA y sobre qué aspectos requieren inevitablemente la intervención humana.

El potencial de la IA no termina en la producción de textos. Otra línea que se presenta con fuerza dentro de los trabajos académicos de los estudiantes de fin de grado de las enseñanzas de Filología es el uso de la IA como herramienta de retroalimentación correctiva. Este volumen incluye ejemplos de experiencias académicas en las que los estudiantes han utilizado herramientas de IA para obtener correcciones automáticas de producciones escritas por aprendices de lenguas y, así, poder evaluar la naturaleza y la calidad de las retroacciones proporcionadas. La ventaja de estas herramientas está clara: ofrecen una respuesta inmediata, objetiva y, potencialmente, pueden ayudar a reducir la carga de trabajo de los profesores de lenguas. Con todo, también plantean un conjunto de cuestiones importantes. ¿Pueden estas herramientas detectar errores complejos de concordancia, estructura discursiva y/o de uso pragmático? ¿Hasta qué punto podemos garantizar que estas herramientas son fiables y adaptadas a diferentes niveles de competencia lingüística y tipologías textuales?

Por último, otro aspecto que presentamos en este volumen es la exploración de la capacidad de las IA para detectar la ironía en la producción lingüística. La ironía es un fenómeno (extra)lingüístico sutil que requiere una comprensión profunda de las convenciones sociales y culturales, así como de las expectativas de los hablantes. La detección de la ironía en un texto representa un reto incluso para los humanos, puesto que a menudo se basa en elementos extralingüísticos, como el tono, la entonación o el conocimiento compartido entre los interlocutores. El uso de bots de conversación capaces de detectar ironía abre una puerta fascinante hacia nuevas aplicaciones de la IA, pero, al mismo tiempo, plantea la pregunta de hasta qué punto estas tecnologías pueden realmente captar los matices de la comunicación humana.

Este cuaderno ofrece una recopilación de experiencias que exploran todos estos aspectos, aportando resultados que serán de gran utilidad para futuros investigadores, estudiantes y docentes interesados en el papel de la IA en estudios sobre la descripción y el análisis lingüísticos, así como en los procesos de enseñanza-aprendizaje de lenguas. El aumento de las propuestas de TFG que exploran estas cuestiones refleja una realidad ineludible: la IA ya no es solo una herramienta, sino un área de estudio con entidad propia, que abre nuevas vías de investigación y que nos obliga a replantear muchos de los supuestos fundamentales sobre cómo abordamos el análisis lingüístico y cómo evaluamos la producción escrita de nuestros alumnos.

Este trabajo colectivo pone de relieve el potencial de la IA no solo como una herramienta de apoyo, sino como un objeto de estudio con el potencial de transformar la forma como investigamos, enseñamos y entendemos la lengua. Los resultados que se presentan aquí son solo un punto de partida en un área que, sin duda, seguirá evolucionando en los próximos años.

1. INTRODUCCIÓN

En las últimas décadas, y más todavía en los últimos años, el impacto creciente de la IA en el procesamiento del lenguaje natural (PLN) ha causado una transformación radical en la forma en que interactuamos con los textos escritos y con las máquinas. La capacidad de la IA para comprender, corregir y generar lenguaje humano ha abierto nuevas perspectivas de investigación y aplicación práctica en muchos campos, como el ámbito educativo, la enseñanza de lenguas (Kristiawan et al., 2024; Almelhes, 2023; Schmidt y Strasser, 2022; Navarro-Carrascosa, Navarro-Carrascosa, 2022; 2024) y el análisis lingüístico (Machado y Ruiz, 2024; Chen et al., 2024; Hromei et al., 2024). En este escenario, deseamos destacar dos líneas de innovación que poseen un alto valor pedagógico y lingüístico: el uso de la IA para ofrecer retroacción autocorrectiva de la producción escrita de aprendices de lenguas y la aplicación para la detección automática de la ironía, un recurso comunicativo muy complejo que históricamente ha sido resistente a la formulación computacional.

> *La incorporación de herramientas de escritura asistida mediante IA es cada vez más habitual en entornos educativos, especialmente en la enseñanza de lenguas y en la adquisición de competencias escritas. Herramientas como Grammarly, Writefull o MyEssai y Large Language Models (LLM) como ChatGPT (OpenAI, 2023) han demostrado una capacidad destacable para identificar errores ortográficos, gramaticales, semánticos y estilísticos, así como para proponer reformulaciones para mejorar el texto, especialmente cuando se trata de lenguas con muchos recursos, como el inglés o el castellano.*

El empleo de estas herramientas presenta claras ventajas en el ámbito educativo, ya que permiten dar una retroacción inmediata y personalizada, cosa que puede resultar especialmente útil en contextos con un alto número de estudiantes o en entornos de autoaprendizaje. Así pues, muchos especialistas destacan la cantidad de retroacción que estas herramientas pueden ofrecer (Guo y Wang, 2023), así como su capacidad de darlo de forma inmediata, lo que posibilita que los aprendices puedan revisar su producción escrita varias veces (Taskiran y Goksel, 2022; Wei et al., 2023). Sin embargo, el uso de esta tecnología tam-

bién plantea dudas sobre la calidad de la retroacción proporcionada, así como cuestiones éticas acerca del uso que los estudiantes hacen de estas herramientas. Asimismo, su aparición abre un debate muy interesante sobre qué papel deben tener en la enseñanza de lenguas y cómo incorporarlas al aula. En particular, es necesario cuestionarse en qué medida estos sistemas contribuyen realmente al aprendizaje o, por el contrario, pueden fomentar una dependencia de la asistencia tecnológica o limitar la creatividad y la flexibilidad expresiva y espontánea de los usuarios.

Otro campo de interés en la intersección entre la IA y la lingüística es la detección de la ironía en textos escritos. Este fenómeno discursivo, que se enmarca en la pragmática, suele implicar una discrepancia entre el significado literal y la intención comunicativa del hablante. Esto hace que detectar automáticamente este rasgo no sea una tarea sencilla. A diferencia de otros tipos de análisis textual, como el análisis de sentimientos o la clasificación temática, la detección de la ironía exige unas habilidades que van más allá de la forma superficial del texto y que incluyen la intencionalidad, la dualidad del significado, junto con una comprensión profunda del contexto, de las expectativas culturales y de las relaciones interpersonales implicadas en la comunicación (Banasik, 2013; Reyes et al., 2013; Wallace et al., 2014).

Los sistemas de detección automática de la ironía han ido evolucionando a la vez que avanzaban los métodos aplicados al PLN. Los primeros sistemas estaban basados en reglas (Tang y Chen 2024; Frenda, 2016) y en enfoques estadísticos que utilizaban diferentes características lingüísticas (Reyes et al., 2013; VanHee et al., 2018a), pero la llegada de las redes neuronales en el área del PLN produjo, como en otras tareas dentro de este campo, un salto cualitativo (Wu et al., 2018; González et al., 2019; Ilić et al., 2018). Trabajos más recientes realizan *fine-tuning*[1] de modelos como BERT y RoBERTa para la detección automática de la ironía (Xaing et al., 2020; Potamias et al., 2020). Estos sistemas necesitan corpus para ser entrenados y evaluados, por lo cual la aparición de estos sistemas ha comportado la compilación y anotación de corpus específicamente diseñados para entrenarlos y evaluarlos, tales

1. En aprendizaje automático, proceso de adaptación de un modelo previamente entrenado para la realización de una tarea específica.

como el corpus Reddit Irony Dataset (Wallace et al., 2014) o el corpus presentado en la tarea de detección de la ironía de SemEval-2018 (Van Hee et al., 2018b). La aparición de LLM y de bots de conversación que permiten que el humano pueda interactuar con la máquina utilizando lenguaje natural ha hecho que la detección automática de la ironía sea aún más necesaria. No obstante, a pesar de los avances técnicos, este sigue siendo un campo donde todavía queda mucho por hacer. La evaluación de cómo estos nuevos enfoques llevan a cabo esta tarea concreta es fundamental a la hora de avanzar en la detección automática de la ironía. Esto plantea la necesidad de un enfoque más interdisciplinar, que combine conocimiento lingüístico, psicológico y computacional, y que considere, también, las implicaciones éticas de la interpretación automatizada de la intención comunicativa.

Esta publicación pretende mostrar tres experiencias enmarcadas en estos dos ámbitos de la aplicación de la IA y del PLN –la retroacción autocorrectiva en textos escritos y la detección de la ironía– como ejemplos de la exploración que se puede hacer de estas tecnologías dentro del contexto de grado.

Dos de estas experiencias analizan el uso de la IA para ofrecer retroacción autocorrectiva de la producción escrita de aprendices de lenguas. En concreto, una se centra en textos escritos por aprendices de inglés como segunda lengua mediante la comparación de la retroacción proporcionada por dos sistemas que utilizan IA, y la otra se enmarca en la enseñanza del catalán como L1 y analiza la retroacción propuesta por ChatGPT. Aunque hay diversas herramientas que pueden ayudar al usuario a corregir y mejorar el texto, como Grammarly, Paperpal o arText, la elección de ChatGPT como una de las herramientas a explorar se debe a que es la herramienta usada de forma más habitual por los estudiantes. Cabe decir que la diferencia entre la retroacción proporcionada para una lengua con muchos recursos disponibles como inglés y la que se proporciona para otra lengua con muchos menos recursos, como el catalán, es muy significativa.

La tercera experiencia que se incluye en esta publicación estudia cómo tres LLM detectan la ironía en textos escritos en inglés, evaluando la capacidad de dichos sistemas para detectarla y clasificarla en una taxonomía propuesta por la autora.

Conviene destacar que el grado de profundidad del análisis varía entre los tres estudios por la naturaleza de los datos y los objetivos de cada trabajo. El dedicado a la detección de ironía adopta un enfoque más cuantitativo, mientras que el de la revisión y corrección de textos en catalán es cualitativo y más detallado, dada la mayor incidencia de errores en esta lengua con menos recursos. Por su parte, el análisis de la retroacción en inglés mantiene también un enfoque cualitativo, pero con menor profundidad, ya que se detectaron menos errores. Estas diferencias metodológicas explican la variación en la extensión y en el nivel de análisis presentado en cada caso.

Estos estudios aspiran, pues, a contribuir a la reflexión sobre el papel de la IA en la didáctica de la lengua y en la investigación sobre el lenguaje, en un momento de cambio acelerado en el cual la colaboración entre disciplinas es esencial.

2. EFECTIVIDAD DE CHATGPT EN LA REVISIÓN Y CORRECCIÓN DE TEXTOS

› **Èlia Vilà Colomé**

1. Introducción

En la actualidad, el progreso acelerado de la tecnología ha conducido a una proliferación sin precedentes de las tecnologías de la información y la comunicación (TIC) y de la IA, que se han integrado cada vez de forma más profunda en nuestra vida cotidiana. Esta expansión ha transformado la forma en que interactuamos con el mundo que nos rodea y afecta desde nuestras rutinas hasta los procesos industriales más complejos.

Las TIC han transformado profundamente la sociedad facilitando la conexión global e impulsando nuevas formas de organización, educación y consumo. En paralelo, la IA ha avanzado significativamente con técnicas como el aprendizaje automático y las redes neuronales, lo que permite a los sistemas informáticos imitar la inteligencia humana, aprender de datos y tomar decisiones. En este contexto emerge ChatGPT, una implementación específica del modelo de lenguaje GPT (*Generative Pre-trained Transformer*), desarrollado por OpenAI. Se vale de una red neuronal de gran escala entrenada para procesar y generar textos de manera coherente. ChatGPT está especialmente diseñado para mantener conversaciones lógicas y relevantes con los usuarios, basándose en el contexto de las interacciones. Este modelo puede usarse en diversas aplicaciones de PLN, destacando por su capacidad de entender y generar textos en distintos idiomas, incluido el catalán. Su capacidad para comprender y producir en catalán ofrece una oportunidad única para su corrección automatizada. El uso de ChatGPT como potencial instrumento de corrección de redacciones en catalán puede contribuir a la optimización del tiempo de los profesores. Podría agilizar el proceso de corrección y permitiría dedicar más tiempo a otras tareas. Además, podría proporcionar retroalimentación inmediata a los usuarios, favoreciendo su competencia lingüística y la capacidad de autocrítica. El objetivo de este trabajo es comprender cómo la IA pue-

de mejorar este proceso y potenciar las habilidades lingüísticas. Si bien existen iniciativas como el proyecto Aina, los recursos disponibles en catalán son limitados e inferiores a los de otros idiomas como el inglés, cosa que evidencia la necesidad de nuevas herramientas de IA para el aprendizaje del catalán.

La propuesta de esta primera experiencia se centra en explorar cómo esta tecnología puede ser utilizada de manera concreta para identificar y evaluar problemas en redacciones académicas en catalán, incluyendo aspectos como la gramática, el léxico, la cohesión y el contenido.

También se busca determinar el *prompt*[2] más adecuado para conseguir una corrección precisa y relevante. Los *prompts* son las instrucciones que se proporcionan a la herramienta de IA para orientar la generación de respuestas. A través de esta exploración se busca comprender hasta qué punto ChatGPT puede ser una herramienta útil para la corrección y mejora de textos escritos en catalán. También se pretende responder a las siguientes preguntas de investigación:

1. ¿Cómo puede ChatGPT ser utilizado eficazmente para evaluar la gramática, el léxico, la cohesión y el contenido en redacciones académicas escritas en catalán?
2. ¿Qué *prompts* son más efectivos para la evaluación de redacciones académicas mediante ChatGPT? ¿Pueden ser adaptados según los distintos niveles educativos?
3. ¿Cuáles son las limitaciones y los retos en el uso de ChatGPT como herramienta de evaluación de redacciones académicas en catalán, y cómo se pueden abordar para mejorar su eficacia y fiabilidad?

2. Metodología

2.1. Herramientas

La herramienta utilizada en este estudio ha sido un modelo de lenguaje desarrollado por OpenAI llamado ChatGPT. Esta herramienta utiliza una arquitectura de aprendizaje automático llamada GPT (*Generative Pre-trained Transformer*) para generar respuestas en texto basadas en

2. Texto o conjunto de datos que un usuario introduce en un sistema de inteligencia artificial para realizarle una petición específica.

las entradas que recibe. El modelo está preentrenado con una gran cantidad de texto de diversas fuentes y corpus lingüísticos con el objetivo de dotarlo de una comprensión profunda del lenguaje humano. En concreto, se ha usado GPT-3.5 (Brown et al., 2020), que es una versión mejorada de la arquitectura GPT que destaca por su capacidad para recordar mensajes anteriores y utilizarlos de contexto en las conversaciones, y para crear respuestas coherentes y relevantes.

Cuando ChatGPT recibe un *prompt*, procesa el texto usando sus capas de atención para entender el contexto y los patrones lingüísticos presentes en la entrada. A continuación, el modelo genera una respuesta basada en dicha comprensión, usando una estrategia probabilística para seleccionar las siguientes palabras de forma coherente y significativa. Esto significa que el modelo asigna probabilidades a las siguientes palabras posibles basándose en el contexto proporcionado y en su conocimiento preexistente. Conforme el modelo avanza en la generación del texto, utiliza el mecanismo de atención para otorgar mayor peso a ciertos elementos de la entrada, y así se asegura que la respuesta generada sea adecuada al contexto, con una estructura coherente y natural.

La eficacia de ChatGPT depende, en gran medida, de los *prompts* que se usen. Para garantizar una evaluación precisa en esta investigación, es preciso encontrar uno que defina los criterios de interés y proporcione un marco de referencia claro. Esto implica diseñar *prompts* pertinentes para el objetivo de la evaluación, abordando los aspectos de la redacción que se quieren evaluar.

2.2. Corpus

Esta investigación se pone a prueba mediante el uso de textos auténticos elaborados por alumnos y profesores, para validar la precisión y utilidad de ChatGPT en este contexto específico. A través de esta metodología, se analizan casos reales y variados de redacciones en catalán. Esta diversidad de muestras facilitará una comprensión más profunda de las capacidades y limitaciones de ChatGPT en el ámbito de la corrección de textos.

El trabajo se basa en el análisis del corpus GRERLI compilado en 1998 y creado en el marco del proyecto internacional «Developing literacy in

different contexts and in different languages». El corpus fue creado por el Grupo de Investigación para el Estudio del Repertorio Lingüístico, con la investigadora Liliana Tolchinsky al frente y Joan Perera como encargado del subcorpus GRERLI-CAT1. Este proyecto tenía como objetivo examinar el repertorio lingüístico de los hablantes de diversas lenguas en una variedad de situaciones comunicativas, explorando las habilidades comunicativas de los participantes, según su nivel educativo.

El corpus GRERLI consta de dos subcorpus: uno en español, llamado GRERLI-CAST1, y uno en catalán, conocido como GRERLI-CAT1. Los textos de esos subcorpus se obtuvieron a través de un mismo procedimiento. Cada uno de los participantes observó un vídeo sin texto ni voz que mostraba situaciones conflictivas en el entorno escolar, y después produjeron cuatro tipos de textos relacionados con este tema: una exposición oral, una exposición escrita, una narración oral y una narración escrita. Se solicitó a los participantes que escribieran una redacción y se destacó la importancia de reflexionar a fondo sobre el tema antes de escribir. Muy probablemente, esta estrategia de utilizar vídeos sin texto ni voz pretendía estimular una respuesta escrita que reflejara las diferentes situaciones de tensión e interacción en el entorno escolar, para reducir las interpretaciones variables que podrían surgir del uso del lenguaje oral o escrito.

El GRERLI-CAT1 incluye textos producidos por 79 hablantes bilingües catalán/castellano de Barcelona (Cataluña), con el catalán como lengua principal. Dentro de este subcorpus hay textos de Educación Primaria, Educación Secundaria, estudiantes universitarios y profesores de Lengua Catalana y Literatura. En total, el GRERLI-CAT1 consiste en 316 textos, 80 de cada nivel, salvo el grupo de Secundaria, que tiene 76. Dentro del corpus disponible, había dos tipos de textos: *expository written* y *narrative written*. Sin embargo, para este estudio solo se han escogido textos del primer tipo, seleccionando veinte en total, cinco de cada nivel educativo. Los dos tipos de textos no presentaban grandes diferencias entre sí, y las respuestas obtenidas por parte de ChatGPT eran similares, lo que hacía irrelevante diferenciar entre ambos.

Esta investigación ha utilizado textos de los cuatro niveles mencionados, para, de este modo, poder comparar cómo varía la corrección dependiendo de los distintos niveles de educación y edades.

Los textos seleccionados (tabla 1) presentan longitudes distintas, debido a la limitación de las opciones disponibles en el corpus. Con todo, se ha intentado garantizar una comparación justa y precisa de las características lingüísticas y discursivas entre los diferentes grupos.

Tabla 1. Corpus de textos utilizados

Nombre	Nivel	Número de palabras
cg01fewa-net	PRIMARIA	107
cg02fewb-net	PRIMARIA	174
cg05fewd-net	PRIMARIA	150
cg08mewc-net	PRIMARIA	90
cg10fewc-net	PRIMARIA	143
cj01fewa-net	ESO	127
cj07mewb-net	ESO	94
cj08fewb-net	ESO	77
cj10mewb-net	ESO	175
cj12mewc-net	ESO	422
cu01fewa-net	UNIVERSIDAD	241
cs07mewb-net	UNIVERSIDAD	294
cu08mewb-net	UNIVERSIDAD	221
cu16fewd-net	UNIVERSIDAD	129
cu17mewd-net	UNIVERSIDAD	433
ct01mewa-net	PROFESORADO	260
ct02mewa-net	PROFESORADO	134
ct08fewd-net	PROFESORADO	147
ct09fewd-net	PROFESORADO	191
ct10mewd-net	PROFESORADO	115

2.3. Procedimiento

El primer paso consistió en la selección de los textos que servirían como corpus para la evaluación de la eficacia de ChatGPT. Con el objetivo de garantizar una recopilación de textos que ofreciera el máximo de retroacción, se escogieron los que presentaban una cantidad significativa de errores gramaticales, ortográficos y de cohesión.

Por tanto, en esta primera etapa del proceso, se dedicó una atención especial a la elección de los textos, poniendo el énfasis en su idoneidad para el objetivo de la investigación. Así, se procuró una selección que proporcionara una muestra representativa del tipo de redacciones que podrían encontrarse en un contexto educativo y que también fuera una base rica en ejemplos concretos que permitiera observar el comportamiento de ChatGPT ante diferentes tipos de errores lingüísticos. Esta selección fue fundamental, puesto que la presencia de errores en los textos permitiría ver realmente la capacidad de detección y corrección de ChatGPT.

El siguiente paso consistió en buscar los *prompts* más adecuados para una evaluación precisa, analizando su eficacia y su adecuación. En el caso específico de corregir textos, un *prompt* adecuado para ChatGPT debería proporcionar contexto suficiente sobre el texto que se quiere corregir. Este *prompt* ha de contener el texto original, indicaciones de las áreas en que se necesita corrección, qué tipo de corrección se desea y cierto contexto para poder situarse. Además, puede incluir una petición explícita para la corrección, por ejemplo, «Corrige los errores ortográficos y gramaticales en el siguiente texto». De esta forma, ChatGPT puede utilizar este *prompt* para comprender el propósito de la tarea y generar una respuesta que ofrezca correcciones coherentes y relevantes para el texto proporcionado.

Se buscaba un *prompt* que proporcionara una corrección eficaz y que se adaptara a los criterios de interés del trabajo. Entre estos criterios, destacaban la corrección gramatical, la adecuación del léxico, la cohesión textual y la coherencia del contenido. Muchos de los *prompts* inicialmente probados no dieron la respuesta esperada, lo cual requirió una continuación de la experimentación hasta encontrar *prompts* que se ajustaran de forma óptima a los objetivos del trabajo. Progresivamente, se refinó el *prompt*, con órdenes como: «Haz un comentario esquemáti-

co sobre la gramática, el léxico, la cohesión y el contenido de ese texto», que empezó a proporcionar respuestas más coherentes. Seguidamente, se especificó que cada texto había sido escrito por una persona de un nivel académico concreto y que se trataba de un ejercicio de clase, esperando que ChatGPT corrigiera como lo haría un profesor.

Cabe decir que al modelo se le puede pedir que adopte el papel de corrector o profesor, pero responde que la información que proporciona puede no ser adecuada o completa, puesto que no tiene la capacidad de analizar o comprender situaciones en tiempo real como lo haría un ser humano. También es importante destacar que las respuestas generadas por ChatGPT son infinitas, es una herramienta en constante desarrollo y las respuestas pueden variar con el tiempo. Las pruebas de este trabajo se llevaron a cabo durante los meses de marzo, abril y mayo de 2024.

Tras detallar el nivel educativo de los autores y el contexto en el que se enmarcaban, se experimentó con diversas formulaciones nuevas de *prompts*. Esta búsqueda culminó en el *prompt* que ha demostrado ser el más eficaz hasta ahora: «Soc alumne d'X/professor i aquest text és un exercici fet a classe. Fes-ne un comentari de tots els errors en la gramàtica, el lèxic, la cohesió i el contingut:». Esta orden proporciona al modelo una orientación precisa sobre los aspectos relevantes de este estudio.

3. Resultados

3.1. Resultados por nivel

La corrección de ChatGPT ha dado como resultado respuestas similares para cada uno de los niveles educativos cuando los consideramos individualmente. Sin embargo, las diferencias se hacen más evidentes al agrupar los niveles en dos grupos: Primaria y ESO, y universidad y profesorado.

3.1.1. Primaria y ESO

Al analizar los textos de los alumnos de Primaria y ESO, puede observarse que ChatGPT detecta y corrige una gran cantidad de errores gramaticales y ortográficos, especialmente los relacionados con la con-

cordancia de número y género. Por ejemplo, corrige a «les dos noies» por «les dues noies», pero no reconoce todos los casos. Además, se han detectado problemas recurrentes, como la inconsistencia a la hora de corregir errores de mayúsculas, minúsculas y acentuación.

En cuanto a las correcciones de textos de ESO, ChatGPT sigue un patrón similar. Se concentra, sobre todo, en la gramática y el léxico, así como en errores de concordancia y errores comunes. Por ejemplo, corrige errores típicos como «hi han» por «hi ha», o también «en el pati» por «al pati». A pesar de que se señalan gran parte de los errores, las correcciones de los textos de Primaria y ESO destacan negativamente por las irregularidades que presentan. A menudo proporciona una solución idéntica al error detectado o una solución errónea. Por ejemplo, en una supuesta corrección sobre el uso de los tiempos verbales, sugiere cambiar «un nen em va ficar el dit a l'ull» por «un nen em va ficar el dit a l'ull», que no modifica la frase original y tampoco contiene ningún error. En otro caso, propone cambiar «t'insulten», que es correcto, por «te insultin», una construcción gramatical incorrecta. Asimismo, no siempre detecta todos los errores presentes en un texto; así, en la frase: «Ademés si ets un marginat», no sugiere corregir «ademés» por «a més» o una expresión equivalente adecuada.

3.1.2. Universidad y profesorado

En los textos de nivel universitario y del profesorado, ChatGPT suele centrarse en recomendar formas diferentes de decir las cosas y ofrece retroacción estilística. Esta pretende mejorar la expresión, la fluidez y la claridad del texto, y no corregir errores concretos. Se han dado casos en los que ChatGPT ha sugerido dividir una frase larga para mejorar su comprensión. También ha propuesto sinónimos para evitar repeticiones, sugiriendo estructuras de frases diferentes para hacer el texto más elegante y cohesionado. Igualmente, ha hecho correcciones adecuadas, como el correcto uso del acento en «què» en lugar de «que» cuando no es una conjunción, la corrección de errores de mayúsculas o la falta de una coma normativa. Sin embargo, pasa por alto otros muchos errores como la ausencia de la primera preposición en la expresión «a poc a poc» o «tant» escrita sin la «t» final del adverbio cuantitativo. Además, no corrige «gimnàsia» por «gimnàstica», «a mida que» por «a mesura que» y «s'enreien» por «se'n reien». No es infalible. Propone sustitucio-

nes inadecuadas, como cambiar «acudits» por «chistes», una corrección descabellada.

En los textos redactados por el profesorado, ChatGPT hace menos correcciones de errores básicos, porque presentan menos faltas. Así que propone la reestructuración de frases para hacerlas más fluidas y precisas. Son cambios sutiles y subjetivos, basados en matices lingüísticos o preferencias.

Otras correcciones inadecuadas e incorrectas han sido la eliminación de «Hi» en «Hi ha hagut», que pasa a «Ha hagut», donde «hi» es una partícula pronominal necesaria para dar sentido completo a la frase y llenar la función de sujeto impersonal. La inconsistencia se ve cuando ChatGPT, en la corrección de otro texto, añade la partícula «hi» cuando es necesaria.

3.2. Resultados por aspecto

3.2.1. Gramática

El apartado de gramática ha sido el más extenso en las correcciones de ChatGPT, con ajustes adecuados sobre todo en Primaria y ESO. Sin embargo, el resultado a menudo ha sido superficial y confuso. Entendemos por *corrección gramatical* el ajustar el uso de los tiempos verbales, las preposiciones y la concordancia, entre otros aspectos. Sin embargo, ChatGPT ha demostrado tener dificultades para delimitar claramente este apartado. En numerosas ocasiones, se incluyen otros comentarios que no tenían relación directa con la gramática estricta, como la puntuación y cohesión.

A pesar de que destaca en la corrección gramatical, debe tenerse en cuenta que algunas de las expresiones señaladas no son estrictamente errores gramaticales y que se han «detectado» errores inventados. Estos errores inventados son de dos clases: en primer lugar, se señala un supuesto error en un verbo correcto y se propone una forma alternativa errónea. Por ejemplo, propone cambiar «Hem pogut veure» por «Hem vist», inventándose un error donde no lo había. En segundo lugar, altera la versión original para aportar soluciones. Por ejemplo, «Els pronoms febles s'han utilitzat incorrectament en alguns casos, com en "és situacions", on hauria de ser "són situacions"», ChatGPT se ha in-

ventado el error y lo ha categorizado mal. En el texto original ya aparecía «són situacions», así que no hay ningún error. En tercer lugar, esta observación nada tiene que ver con los pronombres débiles; si el error fuera verdadero, se trataría de un problema de concordancia de número del verbo, nada que ver con pronombres débiles.

En general, la corrección gramatical no ha sido un fracaso absoluto; de hecho, también se han producido correcciones adecuadas como «hi han» por «hi ha». Este error ha sido identificado de forma consistente por la máquina, como los errores de concordancia, pronombres débiles y pronombres relativos.

3.2.2. Léxico

El análisis del léxico ha sido notablemente inconsistente e insuficiente, puesto que se ha centrado en sugerir mejoras, por ejemplo, proponiendo sinónimos. Sin embargo, muchos de estos sinónimos no han resultado en una solución equivalente al sentido original de la expresión. Además, ha corregido errores de acentuación como léxico, lo cual reafirma la falta de comprensión de las secciones.

En términos de adecuación de registro, ChatGPT ha identificado correctamente que expresiones como «una putada» no son apropiades en un contexto académico, y ha propuesto sinónimos más adecuados. Sin embargo, no ha detectado otros errores, como el uso de «choca» con la «ch» castellana, o la expresión «tenir que». También ha sugerido sustituir «ficar-se amb algú» por «no discutir amb ningú» o «barallar-se», alternativas que no recogen el significado original de la expresión. Una descripción más precisa sería «posar-se amb algú» o «burlar-se'n».

Otro aspecto controvertido ha sido la propuesta de corregir «desequilibrats» por «nerviosos». Socialmente, estos términos no son equivalentes, puesto que el primero tiene una connotación negativa. Esta sugerencia pone de relieve la falta de comprensión de ChatGPT del contexto social y cultural. Además, ChatGPT ha hecho propuestas innecesarias, como sustituir «insolidària» por «manca de solidaritat», «ací» por «aquí» y «deshonesta» por «manca d'honor». En algunos casos, ha modificado palabras del texto original, como en la afirmación: «S'han utilitzat algunes paraules incorrectes o inusuals, com "fet" en lloc de "fetit"». También hay ejemplos en los que no ha entendido el significado concreto

de determinadas palabras. Por ejemplo, al sugerir que «maquineta» se refiere a una «màquina petita» en lugar de a un objeto concreto, o al proponer cambiar el nombre del juego «fet i amagar» por «amagar i buscar», lo cual demuestra una falta de comprensión del contexto.

3.2.3. Cohesión

Las respuestas relacionadas con la cohesión han sido breves, superficiales y generalizadas, limitándose a sugerir una mejora en la conexión entre frases para reforzar la claridad del discurso. En todas las correcciones se han repetido dos observaciones principales: la falta de transiciones claras y la necesidad de introducir conectores para asegurar la fluidez del texto.

Aunque algunos textos presentan indudablemente una estructura fragmentada y una carencia de cohesión, estas observaciones se han aplicado indiscriminadamente, incluso en textos bien redactados. Esto sugiere una falta de criterio específico en la evaluación, puesto que las correcciones no han identificado errores concretos ni han proporcionado propuestas de mejora detalladas.

3.2.4. Contenido

Los comentarios sobre el contenido han seguido la misma línea que los de la cohesión: repetitivos y con explicaciones vagas. ChatGPT no ha proporcionado un análisis demasiado útil, porque se ha limitado a describir el texto sin identificar errores en él. El ejercicio no requería proponer soluciones ni recomendaciones específicas, pero las observaciones podrían haber sido más detalladas y constructivas. Aunque son adecuadas, no aportan una corrección real, sino que señalan posibles mejoras sin profundizar en cómo llevarlas a cabo. Algunos comentarios se centran en el contenido del texto, mientras que otros se enfocan en la cohesión y la estructura. En algunos casos, ChatGPT propone explorar soluciones más a fondo, aunque esto no era necesario para la tarea.

Se hace evidente que ChatGPT no entiende la tarea, puesto que insiste en sugerir el desarrollo de contenido con ejemplos concretos y medidas de mejora, cuando el ejercicio solo requería comentar lo que se veía en el vídeo y las experiencias personales, sin proponer soluciones.

4. Discusión

El análisis tiene como objetivo valorar las correcciones realizadas por ChatGPT. Este proporciona una visión general de los aspectos analizados, incluyendo las discrepancias en las correcciones, la calidad de las sugerencias y las dificultades encontradas en la comprensión de la tarea encargada.

4.1. Impacto del dominio de la lengua y la tarea encargada

ChatGPT aplica un enfoque diferenciado en la corrección de textos en función del nivel educativo. En textos de estudiantes de ESO y Primaria, se centra en errores gramaticales y léxicos, que son más comunes. En cambio, en textos de nivel universitario o del profesorado, tiende a ofrecer recomendaciones estilísticas, ya que los errores lingüísticos son menos frecuentes.

Este enfoque, aun siendo adecuado para las necesidades de cada nivel, no se ajusta a lo que se le había solicitado: limitarse a corregir errores sin ofrecer sugerencias adicionales. Cuando no hay errores, la herramienta debería dejar el texto tal como está. Proporcionar recomendaciones estilísticas, inventarse errores o ampliar temas complica el proceso de revisión y muestra que la herramienta no comprende la tarea asignada. Por ejemplo, ha hecho comentarios como: «No s'aborda completament el tema de com evitar aquests problemes més enllà de l'expressió d'un desig abstracte». ChatGPT considera que sería útil explorar más a fondo otras posibles soluciones, pero este enfoque no es necesario para su corrección. La tarea encomendada consistía únicamente en comentar y corregir los errores presentes en el texto original, sin añadir información ni sugerir ampliaciones temáticas. Este comportamiento de ChatGPT refleja que está programado para dar respuesta completa y detallada en cualquier contexto, incluso cuando no se requiere.

4.2. No detección

Los errores de concordancia de número y género han sido corregidos con gran regularidad, mientras que otros muchos errores no. En algunos casos no se han detectado ni corregido errores como «tenir que» ni

se ha señalado la falta de acentos en «teníem» y «òbviament». También ha habido omisiones en la detección de errores gramaticales, como la expresión «a mida que», que debería ser «a mesura que», ni tampoco la de «choca» con la «ch» castellana. Esta omisión de errores demuestra una limitación significativa en la capacidad de ChatGPT para cumplir plenamente con las necesidades de corrección, lo cual repercute directamente en la calidad del resultado. Esto pone en entredicho la eficacia de la herramienta para una corrección precisa, especialmente cuando se trata de textos de estudiantes. Por consiguiente, es crucial que los usuarios de ChatGPT asuman estas limitaciones y complementen la revisión automática con una revisión humana, en especial en contextos educativos, donde la calidad y la precisión son fundamentales.

4.3. Dificultades con la combinación de pronombres

Se han detectado problemas en la corrección de ChatGPT en lo que atañe a la combinación de pronombres. En algunos casos, la herramienta no ha identificado errores en el uso y colocación de los pronombres, mientras que, en otros, ha efectuado correcciones inadecuadas que no mejoraban el texto y que, de hecho, introducían nuevos errores. Por ejemplo, ChatGPT no ha detectado el error en la frase «me'n enterava». Aquí, el pronombre «me'n» está mal colocado y el verbo «enterava» no es el adecuado. La correcta corrección sería «me n'assabentava», que muestra el verbo correcto y la colocación adecuada del pronombre. De forma similar, la frase «s'enreien» debería ser corregida a «se'n reien, pero ChatGPT no ha señalado este error. En otro ejemplo, ChatGPT ha propuesto una corrección errónea a una frase que ya era correcta. La frase original: «Em vaig adonar que em manava» era clara y bien formada, pero ChatGPT sugirió la corrección: «Em vaig adonar que em estava manant». Esta propuesta introduce un error, pues ignora la necesidad de apostrofar el pronombre «em», como en «m'estava manant».

4.4. Comentarios positivos

Se ha apreciado una carencia de comentarios positivos en las respuestas de ChatGPT en catalán. A pesar de que el usuario ha referido que es un alumno de Primaria, ChatGPT no ha hecho comentarios alentadores ni de ánimo. Esto es especialmente relevante en un contexto educativo, donde el refuerzo positivo puede resultar muy beneficioso

para la motivación de los estudiantes. Los pocos comentarios positivos obtenidos eran de un estilo bastante neutro, como: «És una paraula adient per descriure el comportament violent dels nens» u observaciones como: «La idea central del text és clara» y: «Malgrat la claredat general del missatge, la idea podria ser més precisa». Estos comentarios son funcionales y descriptivos, pero no transmiten unas sugerencias positivas o motivadoras.

Se pueden solicitar este tipo de comentarios de aliento explicitándoselo al *prompt,* pero en este caso no se hizo. En inglés es muy diferente, ya que, sin pedirlo, ChatGPT hace a menudo comentarios positivos. Por ejemplo, al corregir un texto en inglés, se encuentran comentarios como: «You did a great job with your text. I noticed a few mistakes, so let's fix them together».

Estos comentarios muestran una clara diferencia en el enfoque entre las respuestas en inglés y catalán. Mientras que en inglés ChatGPT ofrece un apoyo más positivo y motivador, en catalán falta este componente. Esta diferencia pone de relieve la necesidad de mejorar la aproximación de ChatGPT en catalán a fin de que ofrezca una retroacción más adecuada.

4.5. Corpus y limitaciones

Gran parte de los problemas identificados pueden deberse a la falta de corpus en catalán, lo cual limita el desarrollo de ChatGPT. Una de las causas por las que el modelo presenta dificultades para corregir textos de forma precisa puede ser no tener datos suficientes en esta lengua para aprenderla adecuadamente. Además, ChatGPT puede no estar actualizado con las últimas reformas ortográficas o cambios normativos, puesto que se basa en un conjunto limitado de textos y puede no tener suficiente información. También es posible que no comprenda matices culturales o contextos específicos importantes para una corrección precisa, como referencias culturales, juegos de palabras o expresiones idiomáticas que no son directamente traducibles. Esto también afecta a la comprensión de las diferentes variantes dialectales, los errores tipográficos complejos y las sutilezas semánticas.

Otros problemas con la IA derivan de la falta de sentido común y de conocimiento del mundo por parte de los programas y las máquinas.

Esto puede provocar que corrijan erróneamente cosas en que no es preciso, porque no entienden el contexto, y también pueden pasar por alto errores que sí que requieren corrección. Esta falta de conocimiento del mundo se hace evidente en algunas correcciones; por ejemplo, en un texto universitario, se menciona una «xuleta» escondida «dins una maquineta». ChatGPT sugiere: «L'ús de "maquineta" pot no ser del tot clar. Podries considerar "objecte" o "dispositiu"». Esto revela que no reconoce que se refiere a un sacapuntas para el lápiz y asume, equivocadamente, que se trata de un pequeño dispositivo. También se manifiesta esta carencia de conocimiento cuando sugiere «desequilibrados» como sinónimo de «nerviosos», sin darse cuenta de la connotación negativa que esta palabra tiene en nuestra sociedad. Del mismo modo, sorprende que no comente nada del uso de la palabra «tonta», que podría cambiarse por un sinónimo menos malsonante.

También se han dado casos en los que la máquina ha respondido en castellano, y otras ocasiones en las que ha realizado correcciones en apartados donde no tocaba. Estas incidencias muestran todavía más la necesidad de mejorar la comprensión contextual y lingüística de ChatGPT en catalán, ya que afectan directamente su utilidad y fiabilidad como corrector en dicha lengua. Se puede ver una clara diferencia respecto a cuando se utiliza en inglés, porque, al disponer de un bagaje mucho más amplio, sus respuestas son mucho más acertadas. Esto pone de manifiesto la importancia de ampliar el corpus de datos en catalán, si se quiere conseguir una corrección más precisa y adecuada a las necesidades de los hablantes de la lengua.

5. Implicaciones pedagógicas

La implementación de tecnologías avanzadas como ChatGPT en la educación presenta oportunidades y desafíos significativos. La integración de la IA puede transformar el aprendizaje y la enseñanza, pero exige abordar cuestiones éticas y pedagógicas, como el respeto a la privacidad y la prevención de sesgos. La formación adecuada del profesorado y la personalización de la herramienta son imprescindibles para asegurar un uso efectivo. Antes de generalizar su uso, sería recomendable implementar proyectos piloto para evaluar su efectividad en diversos contextos y obtener datos para mejorar su funcionalidad.

Además, es crucial establecer mecanismos de evaluación continua que analicen el impacto de ChatGPT en el aprendizaje y permitan ajustar su configuración a necesidades específicas. Futuras investigaciones podrían explorar el uso de *prompts* específicos, la aplicación de rúbricas y la comparación con otras herramientas de corrección, así como su efecto en el desarrollo de habilidades lingüísticas. Este enfoque contribuiría a optimizar la integración de la IA en el entorno educativo en beneficio de profesores y alumnado.

6. Conclusiones

El uso de ChatGPT para la corrección de textos en catalán puede ofrecer ventajas significativas, especialmente en lo que se refiere a la velocidad y la precisión en la detección de errores básicos. Sin embargo, su efectividad se ve limitada por la falta de un corpus lo suficientemente amplio y diverso en esta lengua, lo cual puede afectar la capacidad de corrección y la calidad de las sugerencias.

Se han podido observar diversas limitaciones y obstáculos cuando se emplea ChatGPT como corrector de textos en catalán. Actualmente, no es lo suficientemente fiable ni consistente como para ser utilizado como única herramienta de corrección. Con frecuencia pasa por alto errores significativos y puede ofrecer correcciones inadecuadas. Además, la corrección de textos implica la detección de errores gramaticales u ortográficos, junto con una profunda comprensión del contenido, un ámbito en el cual ChatGPT todavía presenta limitaciones. Esta falta de precisión se debe, en parte, a la insuficiencia de un corpus amplio y diverso en catalán, que limita su capacidad para generar correcciones contextualmente adecuadas. Esto pone de manifiesto la necesidad de mejorar el corpus disponible para mejorar la calidad de las correcciones proporcionadas por ChatGPT.

Asimismo, cabe tener en cuenta que ChatGPT tiende a priorizar la complacencia del usuario, lo que puede comprometer la calidad de sus correcciones. Esta característica puede llevar a sugerencias que, aunque parezcan correctas, no siempre son las más adecuadas. Por ello, es crucial que los usuarios apliquen su sentido común y sus conocimientos lingüísticos para evaluar las correcciones propuestas por ChatGPT, en

especial para la detección y corrección de errores más complejos y contextuales o en casos que requieren poseer conocimientos del mundo.

A pesar de estas limitaciones, ChatGPT puede ser una herramienta complementaria muy útil para la revisión de textos, especialmente como primer paso para identificar errores básicos. Aun así, su aportación no es equiparable a la revisión humana y no puede sustituirla por completo. A fin de garantizar una corrección precisa, contextualmente adecuada y de calidad, es imprescindible combinar el uso de ChatGPT con la supervisión y experiencia humanas, puesto que esta colaboración puede mejorar notablemente la eficiencia y la exhaustividad de la revisión de textos en catalán.

De cara a próximos estudios, sería interesante comparar las correcciones proporcionadas por ChatGPT con las facilitadas por otras herramientas, como el corrector ortográfico y gramatical de Softcatalà o el redactor asistido arText. De esta forma se podría valorar qué herramienta es más efectiva y puede serle más útil al usuario.

Agradecimientos

Queremos agradecer la tutorización y guía de la Dra. M. Àngels Massip para la realización de este estudio.

3. RETROACCIÓN GENERADA POR IA EN PRODUCCIÓN ESCRITA EN INGLÉS COMO LENGUA EXTRANJERA DE APRENDICES CON ESPAÑOL COMO L1 A TRAVÉS DE LOS NIVELES DEL MCER: UNA COMPARACIÓN ENTRE MYESSAI Y CHATGPT

› **Júlia Malet Raich**

1. Introducción

En el campo del aprendizaje de lenguas, la IA se utiliza cada vez más, si bien su papel y su eficacia siguen siendo inciertos. La segunda experiencia de este volumen explora la retroacción generada por la IA para comprender mejor cómo esta puede contribuir al aprendizaje de lenguas y ofrecer información sobre sus posibles ventajas y limitaciones.

Este estudio analiza la retroacción de dos herramientas de IA, ChatGPT y MyEssai, sobre 18 textos producidos por aprendices de inglés como lengua extranjera que se preparaban para un examen oficial. Cada alumno redactó un texto transaccional (es decir, un correo electrónico) y un texto creativo (una breve historia). Los participantes se clasificaron en tres niveles de competencia según el Marco Común Europeo de Referencia (MCER): B1, B2 y C1, con tres aprendices representando cada nivel. El objetivo principal del estudio es evaluar y analizar el rendimiento de MyEssai y ChatGPT en su evaluación de los textos de los aprendices.

Para ello, el estudio aborda las siguientes preguntas clave:

1. ¿Cuáles son las características de la retroacción proporcionada por estas dos herramientas de IA?
2. ¿Son estas herramientas de IA capaces de ajustar la retroacción al nivel de competencia del aprendiz?
3. ¿Diferencian las herramientas de IA en función de la tipología del texto?

2. Metodología

2.1. Herramientas de IA: MyEssai y ChatGPT

La primera herramienta de IA utilizada en este estudio es MyEssai, una aplicación diseñada específicamente para proporcionar comentarios sobre los textos escritos por los estudiantes. Utiliza IA para revisar el contenido del texto y proporcionar una retroacción detallada y extensa.

Esta herramienta ha sido escogida porque, a diferencia de otras opciones generales, como ChatGPT, no requiere un *prompt* muy específico para generar la retroalimentación, ya que está programada expresamente para dar retroacción escrita de forma determinada. Sin embargo, para ofrecer una retroacción precisa, es necesario introducir tanto la tipología textual como el enunciado al que responde el texto. MyEssai ofrece diferentes opciones para que la IA se concentre en aspectos concretos, pero este estudio ha trabajado exclusivamente con el enfoque general.

En este estudio se ha trabajado con dos tipologías de texto: *creative writing* (texto narrativo), que encajaba perfectamente con los textos narrativos del corpus, y *cover letter* (carta de presentación), que se eligió como opción más cercana para los correos electrónicos, dado que se trata de textos transaccionales. Además, se facilitaron a la herramienta los enunciados originales a los que los estudiantes habían respondido, a fin de que la retroacción fuera lo más precisa posible.

De esta forma, MyEssai resultaba la mejor opción para este estudio, ya que se ajustaba a las necesidades del corpus y daba una retroalimentación más cuidadosa que otras herramientas de uso más general.

La segunda herramienta es ChatGPT, versión ChatGPT 3.5. A diferencia de MyEssai, ChatGPT no se ha diseñado específicamente para proporcionar retroacción. ChatGPT funciona como un bot de conversación y es un LLM, que significa que ha sido entrenado con un extenso corpus lingüístico. A pesar de que no sea exclusivo para el aprendizaje de lenguas, es capaz de analizar los textos de los estudiantes con cierta profundidad y, por tanto, puede proporcionar retroacción sobre una redacción, si se le pide. Sin embargo, los usuarios deben ser específicos a la hora de dar instrucciones e interaccionar con el bot de conversación, dado que esto afectará la calidad de la retroacción.

2.2. Datos del corpus

Este estudio ha trabajado con el corpus FineDesc, compilado para el proyecto de investigación FineDesc. El corpus, que contiene un total de 800.000 *tokens*,[3] consiste en textos escritos en inglés producidos por aprendices de inglés con español como primera lengua. Concretamente, el estudio ha utilizado 18 redacciones escritas por estos aprendices, que se preparaban para el examen de acreditación en inglés CertAcles. Los textos escritos están divididos en tres niveles de competencia según el MCER: B1, B2 y C1. Por consiguiente, hay seis textos para cada nivel del MCER. Además, los textos también están clasificados según su tipología, por lo que existen tres textos transaccionales (correos electrónicos) y tres textos creativos (textos narrativos) para cada nivel. Así, los escritos están organizados en grupos de tres y cada grupo responde a un enunciado concreto.

2.3. Procedimiento

2.3.1. Elaboración del prompt

MyEssai no necesita un *prompt* muy específico, porque la aplicación ya ha sido creada para proporcionarle retroacción. Aun así, es necesario introducir la tipología del texto, así como el enunciado al cual el texto responde. Como se ha mencionado anteriormente, para este estudio se han seleccionado las opciones de *creative writing* y *cover letter*. La opción de *creative writing* encajaba perfectamente con los textos narrativos del corpus, pero no existía ninguna opción adecuada para los correos electrónicos, por lo que se seleccionó la opción de *cover letter*, ya que era la más similar, al tratarse también de un texto transaccional. Además, se proporcionaron a la IA los enunciados a los que los estudiantes habían respondido para que la retroacción pudiera ser más precisa. MyEssai no permite dar más información (como el nivel MCER) ni pedir retroacción más específica.

En contraste, ChatGPT necesita un *prompt* muy específico y detallado para que la IA analice los textos de la forma deseada. En primer lugar, se indicó a ChatGPT que tenía que revisar una redacción escrita por un

3. Unidad mínima (generalmente a nivel de palabra) que se obtiene al segmentar un texto automáticamente.

aprendiz de inglés con el castellano como L1 y que se preparaba para un examen oficial. Además, se incluyó en el *prompt* el nivel MCER de cada estudiante y el tipo de redacción. Se pidió a ChatGPT que proporcionara comentarios específicos sobre las siguientes áreas: gramática, vocabulario, contenido, organización y resultado de la tarea. Estas áreas se especificaron porque, al pedir simplemente un retorno general, las respuestas eran inconsistentes. Seguidamente, se indicaron las instrucciones que habían seguido los estudiantes para escribir las redacciones. Por último, se proporcionó a la IA la producción escrita que debía analizar. En concreto, el *prompt* utilizado fue el siguiente:

> *I am an EFL learner currently preparing for a high-stakes exam, and my level is a B1 according to the CEFR. Could you give me feedback on grammar, vocabulary, content, organisation and task achievement for my creative writing (100-120 words)? Please, give me feedback on my mistakes, you don't need to provide a revised version. The essay had to begin with the sentence: «He had never called the police before...» This is what I have written: [Text]*

2.3.2. Análisis de la retroacción

Ambas muestras de retroacción generada por IA se compararon para identificar diferencias y semejanzas entre el rendimiento de MyEssai y ChatGPT. En primer lugar, el estudio explora cómo cada una de las herramientas aborda las áreas objeto de revisión (gramática, vocabulario, contenido, organización y resultado de la tarea). El análisis también ofrece información sobre los tipos de retroacción que cada herramienta ofrece con mayor frecuencia. A tal fin, el estudio distingue entre comentarios positivos y correctivos, incluyendo los subtipos de retroacción correctiva: correcciones explícitas, reformulaciones[4] e información metalingüística.[5] Asimismo, se comparan las retroacciones referidas a los textos narrativos con las de los correos electrónicos. Por último, se evalúa la estructura general de la retroacción.

4. Volver a formular algo de otra forma.

5. Uso del lenguaje para hablar del propio lenguaje, analizarlo o explicar su funcionamiento.

3. Resultados y discusión

En la siguiente subsección, se analizará la retroacción proporcionada por ChatGPT y MyEssai en relación con tres áreas principales: lengua, organización y contenido, y resultado de la tarea. A continuación, habrá otra sección que revisará las fortalezas y debilidades generales de MyEssai y ChatGPT, en la cual las herramientas serán analizadas desde una perspectiva más general.

3.1. La retroacción de ChatGPT y MyEssai

3.1.1. Retroacción sobre aspectos lingüísticos

Esta subsección explora la retroacción específica relacionada con los aspectos de lenguaje formal, centrándose en los comentarios de las herramientas sobre la gramática y el vocabulario.

3.1.1.1. Retroacción sobre cuestiones gramaticales

En el contexto de la retroacción correctiva, pueden observarse diferencias y semejanzas claras en la retroacción para la gramática de cada una de las herramientas. Una diferencia notable es que la retroacción de MyEssai es más indirecta que la de ChatGPT. MyEssai utiliza más reformulaciones, con un total de 9 errores, que información metalingüística (6 errores) y correcciones explícitas (6). Por el contrario, ChatGPT tiende a confiar más en las correcciones explícitas, ofreciendo un total de 25 errores, con menos reformulaciones (11) e información metalingüística (6).

Por tanto, ninguna de las herramientas proporciona demasiada información metalingüística y, cuando lo hacen, es, generalmente, breve y no elaborada. Por ejemplo, con la frase: «He was in his flat that night, watching the TV, as he always do», ChatGPT aconseja tener en cuenta la concordancia entre verbo y sujeto: «Use ‘does’ instead of ‘do’ to match the singular subject ‘he’». Para este mismo estudiante, MyEssai destaca el cambio de tiempo verbal, pasando de pasado a presente, y aconseja mantener el pasado: «To maintain clarity, ensure that you consistently use the same tense throughout your story».

ChatGPT suele efectuar más correcciones gramaticales que MyEssai. ChatGPT corrige una media de 4,7 errores gramaticales por aprendiz, mientras que MyEssai solo corrige una media de 2,3. Hay muchas ocasiones en las que MyEssai no corrige errores gramaticales e incluso puede recomendar el uso de otras herramientas como Grammarly para ello. Sin embargo, ChatGPT también pasa por alto errores gramaticales, como en el caso visto en el párrafo anterior, cuando no corrige el cambio de tiempo verbal. Además, hay ocasiones en que ChatGPT resalta errores sin proporcionar correcciones. También puede no especificar errores, como es el caso de un texto narrativo de uno de los aprendices de nivel C1: «sentences are overly complex or convoluted, which affects clarity», pero no menciona qué frases hay que revisar respect a su complejidad.

MyEssai y ChatGPT a menudo corrigen los mismos errores y de forma similar. Por ejemplo, en un texto narrativo de B1, ambas corrigen la frase: «There isn't nobody» a: «There wasn't anybody». Aun así, también hay muchas ocasiones en las que abordan errores distintos. En un mismo texto narrativo de B2, ChatGPT detecta varias inconsistencias de tiempo verbal y sugiere cambiar «make» por «made», «forget» por «forgot» y «see» por «saw». Asimismo, rectifica «that feelings» y propone «those feelings». En cambio, MyEssai solo enmienda los dos primeros errores («make» y «forget») y no menciona el resto. Sin embargo, identifica otro error que ChatGPT no señala: sustituye al plural incorrecto de «woman» por «women». Por ello, ambos corrigen inconsistencias gramaticales, pero no se enfocan en las mismas. Por último, también hay muchos errores que no son mencionados por ninguna de las dos. Por ejemplo, en un texto creativo de B2, ninguna herramienta corrige el error en «a cheap tickets».

Más allá de la retroacción correctiva, también existen diferencias en la retroacción positiva y en las estrategias motivacionales de MyEssai y ChatGPT. MyEssai proporciona constantemente razones para sus correcciones: «Careful attention to spelling and grammar will go a long way in ensuring your readers are immersed in the story and not distracted by these hiccups». Sin embargo, MyEssai no ofrece retroacción positiva específicamente para la gramática. En cambio, ChatGPT normalmente evita explicar el razonamiento detrás de sus correcciones, aunque ocasionalmente proporciona sugerencias constructivas como:

«Simplifying your sentences could improve readability». Además, ChatGPT sí que ofrece retroacción positiva en algunos casos, con comentarios como: «Overall, your grammar is quite good. You demonstrate a good command of sentence structure and verb conjugation».

En cuanto a los distintos niveles de competencia, ambas herramientas muestran poca variación en el tipo de retroacción correctiva. MyEssai dedica menos líneas a la gramática con los aprendices más avanzados, posiblemente a causa de un menor número de errores en niveles más altos. De forma similar, ChatGPT también hace menos correcciones para los aprendices avanzados. En términos de retroacción metalingüística, ChatGPT no proporciona información metalingüística a los estudiantes de nivel C1. En cambio, el enfoque de MyEssai se mantiene constante en todos los niveles, ya que las explicaciones no se adaptan al nivel de competencia de los aprendices. Además, suele ofrecer reformulaciones, aunque la complejidad del lenguaje empleado puede ser demasiado elevada para los aprendices de nivel inferior.

En relación con la retroacción positiva y la motivación, ChatGPT tiende a proporcionar más comentarios positivos en gramática para los aprendices más avanzados. De igual forma, aunque ChatGPT no profundiza demasiado en los aspectos motivacionales, para aprendices de los niveles B2 y C1 hace comentarios alentadores como: «Overall, your grammar is good. You demonstrate a clear understanding of sentence structure and verb conjugation», mientras que, para los aprendices de nivel inferior, a menudo se limita a proporcionar una lista de correcciones. MyEssai no proporciona corrección positiva ni para los estudiantes de nivel inferior ni para los más avanzados. Las estrategias motivacionales utilizadas por MyEssai no se adaptan según el nivel. Para ilustrarlo, comparamos el comentario: «Language and grammar are the threads that stitch your story's fabric», dirigida a un aprendiz de nivel B2, con esta otra dirigida a un aprendiz de nivel C1: «Misspellings and grammatical quirks can momentarily distract readers, so a quick proofreading session is always beneficial».

Por último, MyEssai muestra una mayor variación en las correcciones según la tipología textual en comparación con ChatGPT. En el contexto de los correos electrónicos, MyEssai puede no corregir explícitamente numerosos errores gramaticales, sino centrarse en mantener la forma-

lidad, principalmente mediante reformulaciones. Además, las razones de MyEssai para corregir la gramática varían en función del tipo de texto: en los correos electrónicos, el objetivo es mejorar la claridad y lograr un tono formal, mientras que, en los textos creativos, el objetivo es mantener la inmersión del lector. En cambio, ChatGPT no presenta diferencias en las correcciones gramaticales entre correos electrónicos y textos creativos. Además, las estrategias de corrección positiva y motivación utilizadas para cada tipo de texto son muy similares. Por ejemplo, comparamos este consejo que ChatGPT proporciona para un correo electrónico: «This sentence is quite long and could be divided into two or more sentences for clarity» con este otro dirigido a un texto creativo: «Pay attention to subject-verb agreement and word order to ensure clarity and accuracy in your sentences».

3.1.1.2. Retroacción sobre cuestiones léxicas

En la retroacción correctiva para el léxico, MyEssai y ChatGPT utilizan estrategias diferentes. Aunque MyEssai utiliza correcciones explícitas para abordar problemas de ortografía, también proporciona reformulaciones ampliadas, como la reformulación de la frase en el texto narrativo de un aprendiz de nivel B2. En cambio, ChatGPT utiliza directamente correcciones explícitas, como sustituir «requeriments» por «requested amenities». Además, ChatGPT destaca la importancia de la variedad léxica y recomienda a los aprendices diversificar su vocabulario para mejorar la implicación del lector, por ejemplo, sugiriendo sinónimos para evitar repeticiones. La información metalingüística sobre el vocabulario es muy limitada en la retroacción de ambas herramientas.

Una característica compartida por ChatGPT y MyEssai es que fomentan el uso de un lenguaje descriptivo. Esto puede observarse en la retroacción en el texto narrativo de un aprendiz de nivel B2, a quien ambas herramientas recomiendan incluir más detalles en su narración. ChatGPT explica: «using more descriptive language to paint a vivid picture of the event and your experiences». Del mismo modo, MyEssai aconseja: «Include sensory details such as the smell of burning rubber, the feel of the vibrations from the roaring engines, or the sight of dusk settling over the rally».

Pese a que la retroacción positiva sobre vocabulario no es muy frecuente, ambas plataformas ofrecen ocasionalmente comentarios alentadores. ChatGPT felicita a un aprendiz: «Your choice of vocabulary is generally appropriate and helps to convey your experience vividly». MyEssai destaca un punto fuerte en la escritura de un aprendiz: «You've chosen a conversational tone, which makes the story approachable, and that's a strength to build on». Sin embargo, este tipo de comentarios de ánimo son escasos en la retroacción de ambas herramientas. En cambio, en lo que se refiere a la motivación, las herramientas suelen explicar por qué es importante seguir sus recomendaciones. En una ocasión, MyEssai argumenta: «Descriptive language is a powerful tool to transport readers into the scene». De manera similar, ChatGPT afirma: «Use a variety of vocabulary to make your writing more engaging».

Existen algunas diferencias en el enfoque dependiendo del nivel de competencia del aprendiz entre MyEssai y ChatGPT. MyEssai no adapta la retroacción al nivel del aprendiz; de forma similar a su retroacción sobre gramática, las reformulaciones que proporciona a menudo incluyen vocabulario que puede ser complicado para los aprendices de nivel inferior. En cambio, ChatGPT adapta su retroacción en cierto grado. Para los aprendices de nivel inferior, ChatGPT se centra en corregir errores de ortografía y selección de palabras, por lo que las correcciones que proporciona son más básicas, tales como: «It would be more appropriate to refer to it as a 'post' or 'listing' rather than a 'publication'». En cambio, para los aprendices más avanzados, ofrece sugerencias orientadas a mejorar la variedad léxica: «Your vocabulary is appropriate for the context, but consider varying your word choices to avoid repetition. For example, instead of using 'materials' multiple times, you could use synonyms like 'resources' or 'content'».

ChatGPT y MyEssai parecen adaptar las observaciones sobre vocabulario de acuerdo con el tipo de texto. Ambas herramientas ponen el énfasis en el lenguaje descriptivo en los textos narrativos para mejorar la inmersión en la historia, mientras que priorizan la formalidad y el profesionalismo en los correos electrónicos. Por ejemplo, en un correo electrónico, ChatGPT sugiere cambiar: «I could consider this a scam» por: «I believe this situation may constitute fraudulent misrepresentation», consiguiendo un tono mucho más formal. Sin embargo, en los textos narrativos, aconseja utilizar un lenguaje más descriptivo para

crear una imagen más clara para el lector. De forma similar, MyEssai sugiere moderar el lenguaje contundente en los correos electrónicos y reformular ciertos términos para transmitir más diplomacia. Para los textos narrativos, MyEssai efectúa recomendaciones más enfocadas en las sensaciones transmitidas por el lenguaje: «More vivid and sensorial language could enhance the atmosphere».

3.1.2. Retroacción sobre la organización del texto

Tanto MyEssai como ChatGPT realizan sugerencias similares para organizar los textos, destacando la importancia de la claridad y la coherencia. Ambas herramientas recomiendan con frecuencia dividir los textos en más párrafos para mejorar su legibilidad. Por ejemplo, MyEssai sugiere escribir párrafos más cortos: «White space in writing allows the reader to absorbe thought before moving on to the next, which can be quite effective when dealing with abstract concepts». De manera similar, ChatGPT afirma: «The writing lacks clear paragraph breaks, which can make it challenging to read. Organize your ideas into paragraphs to improve clarity and coherence». De hecho, ambas herramientas aconsejan ubicar cada idea en un párrafo separado. ChatGPT sugiere: «Each paragraph could focus on a specific aspect of how music breaks down boundaries or connects people, allowing for a smoother flow of ideas…». MyEssai recomienda al aprendiz: «Avoid jumping from one topic to another without a smooth transition. To improve flow, group related questions together and ensure each paragraph has a clear focal point».

Además, ambas herramientas abordan la complejidad de las oraciones. MyEssai señala que, si bien presentar ideas complejas es positivo, las oraciones densas pueden dificultar la comprensión del significado. Por eso, recomienda dividirlas en estructuras más simples para ganar claridad. ChatGPT coincide en este punto y aconseja dividir las oraciones largas en otras más cortas a fin de mejorar la legibilidad. Ambas herramientas también sugieren a varios aprendices empezar los correos electrónicos con una frase introductoria para ofrecer contexto.

Las retroacciones proporcionadas por MyEssai y ChatGPT presentan tanto coincidencias como diferencias en la retroacción para un mismo aprendiz. Por ejemplo, en un correo electrónico de nivel B1, las herra-

mientas coinciden en la importancia de una estructura clara y ofrecen retroacción positiva antes de plantear sugerencias de mejora. Las dos hacen hincapié en la necesidad de reorganizar los párrafos agrupando preguntas similares para facilitar la comprensión, pero MyEssai es más detallado en su explicación y proporciona un ejemplo concreto de cómo estructurar el texto. Sin embargo, en otras ocasiones se pueden observar discrepancias, como es el caso del correo electrónico del aprendiz 3 del mismo nivel. Mientras que ChatGPT sugiere únicamente dividir las oraciones más largas para mejorar la legibilidad, MyEssai propone una reorganización más profunda, incluyendo la necesidad de presentarse adecuadamente y explicar la motivación antes de formular preguntas. Por tanto, se contradicen, dado que ChatGPT considera que la organización de los párrafos es adecuada, mientras que MyEssai apuesta por cambiarlos radicalmente.

Asimismo, con algunos textos, MyEssai no ofrece ningún tipo de retroacción sobre la organización; por ejemplo, en el caso del texto narrativo del aprendiz 1 de C1. Se trata de un aprendiz avanzado, por lo que podría pensarse que MyEssai no ofrece consejos de organización porque considera que en este texto no son necesarios. Sin embargo, esta suposición contrasta con la retroacción de ChatGPT para el mismo texto del aprendiz: «The essay lacks clear organisation and transitions between ideas. Consider structuring it into paragraphs to improve readability and coherence». Otra posible explicación a la falta de retroacción sobre la organización por parte de MyEssai es que la IA decide que el aprendiz ha de priorizar otras áreas. En cualquier caso, no es fácil discernir la razón por la cual MyEssai no hace ningún comentario sobre la organización.

Tanto MyEssai como ChatGPT ofrecen motivación y retroacción positiva, pero las aproximaciones y frecuencia varían. Como ocurre con otras áreas, MyEssai se centra en explicar las razones detrás de las sugerencias de organización: «Structuring your complaints like this gives the letter recipient a quick, clear understanding of your concerns». ChatGPT no se extiende tanto en las razones, aunque puede mencionarlas brevemente en algunas ocasiones: «Consider structuring it into paragraphs to improve readability and coherence». Respecto a la retroacción positiva, MyEssai tiende a ofrecer menos, pero, cuando lo hace, incluye comentarios detallados y específicos, como: «The introduction is effec-

tive in establishing the universal resonance of music. The phrase 'It has the power to give us peace and the strength to make us jump' is striking and sets the duality of music's impact beautifully». En cambio, ChatGPT casi siempre proporciona retroacción positiva: «The organisation of ideas is clear, with a beginning, middle, and end».

No existen diferencias significativas en los niveles de competencia entre la retroacción proporcionada por ambas herramientas. Para los aprendices de B1 y B2, MyEssai se preocupa por ofrecer un orden más natural y lógico, aconsejando a los aprendices que diferencien claramente entre introducción, cuerpo y conclusión. También recomienda habitualmente el uso de oraciones más cortas y párrafos más cortos. Con la mayoría de los aprendices C1 no se extiende demasiado en aspectos de organización, pero los escasos comentarios que hace siguen la misma línea que los dirigidos a niveles inferiores: «In rewriting, you might want to work with shorter paragraphs to create a more impactful pace». De forma similar, en cuanto a la retroacción correctiva, ChatGPT tiende a dar la recomendación de organizar el texto en párrafos más cortos y agrupar ideas similares. Sin embargo, solo ofrece retroacción positiva a los aprendices de B2 y C1, con comentarios como: «The email is reasonably well-organized, with each paragraph focusing on a different aspect of the complaint».

MyEssai adapta la retroacción correctiva al tipo de texto. En los correos electrónicos, la retroacción tiene el objetivo de hacerlos más comprensibles, sugiriendo maneras lógicas y efectivas de organizar el texto: «Avoid jumping from one topic to another without a smooth transition. To improve flow, group related questions together and ensure each paragraph has a clear focal point». La correctiva retroacción que MyEssai ofrece para los textos narrativos no solo tiene el objetivo de hacer la historia más clara, sino también más inmersiva y fluida. Por ejemplo, ofrece este consejo: «Regarding the structure, an engaging short story often has a beginning that hooks the reader, a middle that builds upon that interest, and an end that leaves them thinking».

Por el contrario, la retroacción correctiva que ChatGPT da para cada tipo de texto no presenta contrastes notables. Por ejemplo, comparamos estos consejos casi idénticos: «Consider breaking the text into smaller paragraphs for better readability» y: «Consider breaking the

text into smaller paragraphs to improve readability and flow». Sin embargo, la retroacción positiva es más diferenciada, ya que reconoce los objetivos propios de la tipología textual. Por un lado, para un correo afirma: «The email is well-organized, with each paragraph focusing on a different aspect of the problem». Así, destaca que el texto es claro y comprensible. Por otra parte, en los textos narrativos da prioridad a la fluidez: «The story flows logically from your anticipation before the performance to the climax of experiencing the music and emotions during the opera». También puede decirse que ChatGPT ajusta hasta un punto la retroacción organizativa según el tipo de texto.

3.1.3. Retroacción sobre el contenido y el resultado de la tarea

MyEssai siempre proporciona sugerencias extensas relacionadas con el contenido del texto. Por ejemplo, para el texto narrativo de un aprendiz de nivel B2, MyEssai realiza brinda recomendaciones de este tipo, como explorar las emociones del escritor, proporcionar más detalles sobre los personajes y expandir la conclusión, dedicando un párrafo completo a cada una. En cambio, ChatGPT tiende a ofrecer menos sugerencias o a solo incluir retroacción positiva. Para ilustrarlo, para el aprendiz mencionado anteriormente, ChatGPT hace dos comentarios positivos: «You effectively convey the excitement and anticipation of attending the opera for the first time» y: «You describe the impact of the music and performances on your emotions well, which enhances the storytelling». Esta comparación indica que MyEssai proporciona consejos detallados y específicos sobre el contexto, mientras que la retroacción de ChatGPT no entra en recomendaciones detalladas para mejorar.

En relación con la consecución de la tarea, ambas herramientas de IA a veces ofrecen retroacciones contradictorias. Por ejemplo, en lo que se refiere al texto narrativo de un aprendiz de C1 que no responde correctamente al enunciado, ChatGPT afirma: «The essay effectively addresses the prompt by discussing how music transcends boundaries and connects people emotionally». Sin embargo, MyEssai sí que identifica que el aprendiz se desvía de la labor: «The quote you are reacting to speaks specifically to music's ability to communicate in ways that surpass verbal and physical expression». Además, MyEssai incluye una sugerencia para ajustar la escritura más estrechamente a la tarea: «Consider focusing on a narrative that directly illustrates this unique

ability of music to transcend those boundaries, instead of speaking broadly about its various applications». En cambio, con uno de los textos narrativos de nivel B1, solo ChatGPT identifica que la tarea no se ajusta bien al texto: «While the story begins with the given sentence, it deviates from the prompt's context, focusing more on the protagonist's experience with BTS tickets rather than an incident requiring a call to the police». En todo caso, no se da ningún consejo para corregir este problema.

Cuando se analiza la retroacción sobre el contenido para cada nivel de competencia, se aprecian diferentes patrones entre ChatGPT y MyEssai. ChatGPT proporciona tanto retroacción correctiva como positiva para los aprendices B1, pero para los niveles B2 y C1 solo ofrece retroacción positiva. En contraste, MyEssai muestra pocas diferencias en su enfoque de retroacción según los niveles. Por ejemplo, en el texto narrativo de un solo aprendiz de nivel B1, MyEssai sugiere dar más detalles, desarrollar a los personajes, mejorar el desarrollo de la acción para evitar un final brusco, etc. Además, cada recomendación ocupa un párrafo. Así, las numerosas sugerencias y su extensión prolongada pueden resultar abrumadoras para un aprendiz de ese nivel. Esto indica que, mientras que ChatGPT ajusta, hasta cierto punto, la complejidad de la retroacción sobre el contenido según el nivel del aprendiz, MyEssai proporciona constantemente retroacción detallada y completa, independientemente del nivel de competencia del aprendiz.

En cuanto a la tipología de texto, MyEssai parece adaptar la retroacción sobre el contenido para proporcionar las correcciones y sugerencias más adecuadas. En los textos narrativos, pone el énfasis en la creación de una narrativa más atractiva y evocadora, proporcionando retroacción detallada y específica sobre el contexto, animando a los aprendices a transmitir vívidamente las emociones y las experiencias. En la redacción de correos electrónicos, MyEssai se concentra en el profesionalismo y la formalidad, manteniendo la claridad y la eficacia: «It's good to make the personal impact clear, but it might help to stick to the most pertinent details to keep the letter concise».

De forma similar, ChatGPT ofrece consejos distintos según el tipo de texto. Por ejemplo, para un texto narrativo comenta: «There's room for further development of the character's emotions and reactions to the

unfolding events. Adding more depth to the protagonist's thoughts and feelings can enhance the story's impact». También persigue que las narraciones sean coherentes y claras: «It's not clear why the protagonist's friend didn't respond or why the police were called». En la redacción de correos electrónicos, ChatGPT no hace tanto hincapié en la formalidad como MyEssai, sino que se centra en la claridad en la comunicación: «Some questions are unclear or could be more specific. For example, you could ask about specific challenges Stephanie faced while living alone rather than just asking what she learned».

3.2. Fortalezas y debilidades generales de ChatGPT y MyEssai

Hay diferencias significativas en la forma como se han diseñado MyEssai y ChatGPT, y, por consiguiente, los comentarios que pueden proporcionar son distintos. Una ventaja importante de ChatGPT es su capacidad de ser personalizado. Los usuarios pueden solicitar comentarios específicos o adaptarlos a sus características o necesidades individuales. Sin embargo, esto requiere un *prompt* preciso, ya que las consultas ambiguas pueden llevar a respuestas inexactas. Además, sus comentarios tienden a ser más breves y, en ocasiones, superficiales, como se demuestra viendo que ofrece las mismas sugerencias exactas a diferentes aprendices. En cambio, MyEssai ofrece comentarios extensos y específicos del contenido, pero carece de la capacidad de personalización de ChatGPT. No puede ajustarse al nivel de competencia del usuario ni centrarse en áreas específicas.

La forma en que las herramientas de IA proporcionan comentarios es distinta, con cada herramienta utilizando su propio método. MyEssai, normalmente, tiene un tono más amable: «Your plot takes an unexpected turn, which is great for surprise, but the resolution feels a bit abrupt. Perhaps you could build up to the twist more gradually». Así pues, intenta hacer comentarios positivos para suavizar el impacto de la crítica. En cambio, ChatGPT a veces es muy directo: «The writing lacks clear paragraph breaks, which affects readability and organisation». La crítica se presenta sin atenuación alguna. Además, a menudo destaca aspectos negativos sin sugerir formas de mejorar. Por ejemplo, le dice a un aprendiz de nivel B1: «The sequence of events could be clearer and more logically connected», pero no aporta consejos sobre cómo hacerlo.

MyEssai siempre comienza y termina los comentarios motivando al aprendiz. Empieza alabando al aprendiz con comentarios como: «Firstly, I want to commend you for the professional tone of your cover letter. You've presented your complaint in a measured and courteous manner, and that is definitely the right approach when addressing a contentious issue». Además, la conclusión de los comentarios intenta alentar al aprendiz a que siga progresando: «Keep writing, as your voice is as deserving of an audience as the opera singers who've inspired you. Your passion for the subject is the melody of your prose – let it sing with the same grace and strength you admired on stage». Como se muestra en este fragmento, los comentarios motivacionales no parecen vacíos, puesto que están relacionados con el contexto del texto. Además, MyEssai a menudo incluye un resumen de todos los comentarios en la conclusión, proporcionando una visión general completa para el aprendiz.

ChatGPT no es tan consistente en la provisión de comentarios motivacionales al principio, por el hecho de que solo ofrece halagos para los aprendices de B2 y C1, y suelen ser comentarios más breves: «Your short story effectively captures the excitement and passion you have for the WRC-Catalonia event». En la conclusión, ofrece declaraciones alentadoras a todos los aprendices, que suelen ser más largas: «Overall, you've done a commendable job of crafting a short story about your experience at the opera. Paying attention to minor grammar errors and refining your vocabulary choices can help enhance the clarity and impact of your writing. Keep up the good work!». A pesar de que no sean tan personalizados como los de MyEssai, los comentarios también son individuales para cada aprendiz, puesto que se refieren a su texto. En lugar de resumir los puntos importantes, ChatGPT tiende a mencionar las áreas que cree que necesitan más mejora: «Paying attention to grammar and organisation can help enhance the clarity and impact of your writing».

4. Implicaciones pedagógicas

Las implicaciones pedagógicas de este trabajo están relacionadas con el papel que MyEssai y ChatGPT podrían desempeñar en el ámbito de

la enseñanza-aprendizaje de lenguas. El estudio apunta a que los estudiantes podrían utilizar estas herramientas para facilitar el trabajo autónomo, dado que proporcionan comentarios útiles. Además, son capaces, hasta cierto punto, de adaptar el tipo de retroalimentación según los errores que cometen los aprendices, además de incluir comentarios positivos y motivacionales. Aun así, algunos problemas detectados en el análisis, como el ajuste limitado al nivel del aprendiz y la falta de información metalingüística, indican la necesidad persistente de la intervención humana para ayudar a los aprendices a comprender y asimilar la retroacción.

5. Conclusiones

Hay similitudes interesantes, así como diferencias, entre la retroacción correctiva proporcionada por MyEssai y ChatGPT. Por lo que respecta al lenguaje, la retroacción de MyEssai es más indirecta que la de ChatGPT, que tiende a utilizar muchas correcciones explícitas. Ninguna de las dos herramientas proporciona demasiada información metalingüística. ChatGPT corrige 89 errores formales, mientras que MyEssai solo 50. Esto, junto con el hecho de que recomienda el uso de Grammarly en diversas ocasiones, indica que el objetivo principal de MyEssai no es corregir estos errores, sino que se preocupa más por el contenido. De hecho, la mayor parte de la retroacción de MyEssai se centra en el contenido del texto, dedicando varios párrafos que incluyen observaciones estrechamente relacionadas con el texto. Por el contrario, los comentarios de ChatGPT sobre el contenido a veces parecen superficiales. En todo caso, el rendimiento varía en e resultado de la tarea: a veces es ChatGPT el que detecta desviaciones del enunciado, mientras que otras veces solo lo hace MyEssai. En términos de organización, MyEssai tiende a ser más específico, ofreciendo ejemplos de cómo estructurar el texto, mientras que ChatGPT suele dar consejos breves.

Respecto a la retroacción positiva y la motivación, ChatGPT proporciona retroacción positiva en más áreas que MyEssai, y en ocasiones excluye la retroacción correctiva. En cambio, MyEssai busca establecer una conexión amigable con el aprendiz antes de proporcionar retroacción correctiva, mientras que ChatGPT puede ser bastante directo. Además, MyEssai hace un mayor esfuerzo para ayudar al aprendiz a entender

la importancia de mejorar los aspectos que sugiere. Si bien ChatGPT incluye comentarios similares, estos son menos frecuentes en comparación con los de MyEssai. Ambas herramientas de IA terminan sistemáticamente los comentarios animando a los aprendices a seguir progresando, un elemento crucial para su motivación.

Ninguna de las dos herramientas parece capaz de ajustar la retroacción al nivel de competencia del aprendiz. Sin embargo, ChatGPT es el que muestra más diferencias en la retroacción en función del nivel del aprendiz, mientras que MyEssai no hace distinción alguna y la extensión de su retroacción puede ser contraproducente para aprendices de nivel inicial. Por tanto, la retroacción detallada de MyEssai puede ser más útil para los aprendices avanzados, que sabrán mejor cómo aplicar sus correcciones a los textos. Sin embargo, la simplicidad y la claridad de ChatGPT pueden ser más útiles en los niveles iniciales.

En cambio, MyEssai adapta la retroacción de forma más efectiva al tipo de texto. Por ejemplo, en los correos electrónicos, se centra especialmente en aspectos de formalidad y en escribir de forma eficiente para que el destinatario pueda entender claramente las intenciones del autor. En la escritura creativa, ofrece numerosas sugerencias para que las historias sean más atractivas y evocadoras. ChatGPT también ajusta la retroacción sugiriendo historias más cautivadoras y correos electrónicos más comprensibles, pero se adapta menos, puesto que sus recomendaciones entre tipologías textuales no difieren tanto como las de MyEssai.

En conclusión, aunque MyEssai y ChatGPT comparten algunas características, su rendimiento varía significativamente. Por consiguiente, la elección de la herramienta debería depender de las circunstancias individuales y de las necesidades específicas del aprendiz. Todavía hacen falta más investigaciones para explorar el impacto que estas herramientas pueden tener en la producción escrita de los aprendices. Sin embargo, el estudio actual ha demostrado que, si bien ambas herramientas de IA presentan puntos fuertes, también tienen problemas, por lo que es necesario complementarlas con la retroacción proporcionada por los docentes. Como trabajo futuro, podría ampliarse el estudio comparando las correcciones proporcionadas por otras herramientas (por ejemplo, Grammarly o Paperpal) u otros LLM, tales como Claude o Gemini.

4. EVALUACIÓN DE LA CAPACIDAD DE LOS BOTS DE CONVERSACIÓN CON IA EN LA DETECCIÓN DE IRONÍA: UN ESTUDIO DE CHATGPT, COPILOT Y GEMINI

› **Ionela Valentina Vasilovici**

1. Introducción

La ironía es un recurso lingüístico importante en relación con el habla humana, puesto que se utiliza para transmitir ideas complejas, emociones y sentimientos, entre otras cosas. Además, forma parte de la creatividad del lenguaje humano y cumple diversas funciones. Así pues, a medida que avanza la IA, va aumentando el interés por hacer que las máquinas capturen el lenguaje no literal de los humanos y vayan más allá de un simple conjunto de palabras, ya que la detección de la ironía tiene muchas aplicaciones (por ejemplo, análisis de sentimientos o de opiniones). Los investigadores en IA tienen como objetivo mejorar la comunicación entre el ser humano y la máquina, siendo la ironía verbal un elemento clave en el lenguaje no literal y en el campo de la pragmática.

En este contexto, la aparición de ChatGPT y otros bots de conversación con IA o LLM como Copilot (antes Bing) y Gemini (antes Bard) representa un avance significativo, dado que pueden entender un *prompt* y proporcionar una respuesta creativa. En otras palabras, con la aparición de estos bots de conversación, se abre un nuevo abanico de posibilidades para explorar su comprensión de la producción (escrita y oral) humana y su capacidad para detectar ironía y clasificar funciones comunicativas. Además, la comprensión de la ironía verbal permite a los bots tener una idea más clara de lo que se solicita. Por tanto, es interesante analizar su potencial y sus limitaciones para contribuir al campo de las tecnologías del lenguaje y a su continuo desarrollo. La última propuesta presentada en este volumen se adentra en la detección de la ironía para evaluar hasta qué punto la IA puede comprender y detectar fenómenos lingüísticos complejos. Más específicamente, el objetivo del estudio es poner a prueba la capacidad de dichos bots de conversación para entender cómo abarcan un recurso lingüístico complejo como la ironía y cómo clasifican las funciones comunicativas. Para explorar

este tema, este estudio aborda las siguientes preguntas de investigación:

1. ¿Cuáles son los puntos fuertes y las limitaciones de los bots de conversación con IA como ChatGPT, Microsoft Copilot y Google Gemini respecto a la detección de la ironía?
2. ¿Cuál de ellos es más preciso? ¿Cuáles son las diferencias?
3. ¿Qué tipo de *prompt* (*zero-shot* o *few-shot*) funciona mejor con respecto a la clasificación de la ironía?

2. Metodología

2.1. Diseño del guion para la detección de la ironía

Con el fin de explorar si ChatGPT, Copilot y Gemini son capaces de detectar ironía, y para poder valorar sus resultados, se ha creado un guion que consiste en una conversación entre dos personajes:

> Personaje 1 (Ally): Su función principal es expresar la mayoría de los enunciados irónicos.

> Personaje 2 (Daniel): Su función principal es dar respuesta a estas expresiones y permitir que la conversación continúe.

Los hablantes inician una conversación mientras esperan su tren y el Personaje 1 se queja constantemente de su trabajo y de su compañera de piso, utilizando muchas expresiones irónicas. El guion incluye diecisiete ejemplos de ironía que pueden clasificarse según su recurso retórico (hipérbole, metáfora, exclamación, pregunta retórica y símil), ya que el objetivo es contar con una variedad de expresiones irónicas. Además, las categorías también se utilizarán en la interacción con los bots de conversación. Se han incluido tres hipérbolas, cuatro metáforas, cuatro exclamaciones, tres preguntas retóricas y tres símiles. El Personaje 1 es el que expresa la mayoría de los ejemplos de ironía (15). En cambio, el Personaje 2 solo produce dos ejemplos para explorar si existe alguna diferencia relevante en cuanto a la detección de la ironía. La tabla 1 contiene todos los ejemplos de ironía con el correspondiente

recurso retórico, en el mismo orden en que aparecen en el guion. Sin embargo, también incluye una columna con otras categorías que serían aceptadas, puesto que algunas pueden superponerse.

Tabla 1. Expresiones irónicas del guion y sus categorías correspondientes

Expresión irónica	Categoría principal	Otras categorías aceptadas
1- It's as cold as ice today, huh?	Símil	Hipérbole
2- My job is a dream!	Exclamación	Hipérbole, metáfora
3- You can imagine that my boss is such an angel.	Metáfora	Hipérbole
4- Do you also have a boss who is as happy as a clam?	Símil	-
5- As his voice was music to my ears...	Metáfora	-
6- I love working here for 12 hours in a row!	Exclamación	Hipérbole
7- I have just been waiting here for 5 hours.	Hipérbole	-
8- I can see that calmness is your friend.	Metáfora	-
9- This is the best week of my life!	Exclamación	Hipérbole
10-...no one dares to complain about the best boss of all time.	Hipérbole	-
11- She is always as busy as a bee.	Símil	-
12- Aren't you tired of studying all day?	Pregunta retórica	-
13- Who wouldn't want to read twenty romance books in a month?	Pregunta retórica	-
14-...how could I be so absent-minded as to forget a book that I never picked up?	Pregunta retórica	-
15- I have done this a million times!	Exclamación	Hipérbole
16- Does it have something to do with the super-organised flatmate?	Hipérbole	-
17-You are the worst kind of monster.	Metáfora	Hipérbole

Estos ejemplos de ironía se insertan artificialmente en la conversación para que el bot los detecte. De cara a facilitar la interpretación, incluyen fuertes contrastes en el entorno. Por ejemplo, existe una gran contradicción en esta afirmación: «¡She is always as busy as a bee: chilling

on the couch all day!». Así pues, el contexto y los contrastes son clave en este texto. Además, la conversación incluye ejemplos artificiales de ironía en lugar de una conversación natural, puesto que debe incluir una cantidad representativa de expresiones irónicas de cada categoría (tres o cuatro ejemplos de ironía de cinco categorías en una conversación natural no es realista). Por último, el tono humorístico general contribuye a la comprensión de los enunciados irónicos.

2.2. Aplicación del guion para la detección de la ironía

Antes de profundizar en la interacción con los bots de conversación, se han de tener en cuenta las características de cada uno y las propiedades que puedan ser relevantes, como se muestra en la tabla 2.

Tabla 2. Bots de conversación utilizados en este estudio y sus características

Bot de conversación con IA[6]	LLM	Precio	Inicio de sesión	Propiedades	Fechas de acceso
ChatGPT	GPT-3.5	Gratuito (versión de pago disponible basada en GPT-4)	Creando un cuenta en sitio web	-	4/5/2024 - 21/5/2024 (actualización el 16 de mayo)
Copilot	GPT-4	Gratuito	Mediante una cuenta de Microsoft	Estilo de conversación preciso	4/5/2024 - 20/5/2024
Gemini	PALM-2	Gratuito (versión de pago disponible llamada Gemini Advanced)	Mediante una cuenta de Google	-	4/5/2024 - 20/5/2024 (actualización el 14 de mayo)

Los datos de la tabla 2 indican que tanto GPT-4 como PaLM-2 son tecnologías más avanzadas que GPT-3.5. Además, Copilot ofrece tres estilos de conversación distintos: creativo, equilibrado y preciso. El estilo creativo es para iniciar un chat original e imaginativo, el estilo equilibrado es para chats cotidianos y el estilo preciso es para iniciar un chat conciso (útil para la búsqueda de hechos). A efectos de este estudio, y en

6. Inspirada en la tabla de la búsqueda de Morreel et al. (2024) https://doi.org/10.1371/journal.pdig.0000349.t001

relación con la detección de ironía, el estilo conversacional preciso es el que mejor funciona.

Por lo que respecta a la interacción, la conversación con los bots consiste en dos pasos y dos tipos de preguntas, como se ilustra en la tabla 3.

Tabla 3. Tipos de técnicas y *prompts* utilizados en la fase de interacción

	***Prompt* 1: Identificación**	***Prompt* 2: Clasificación**
Zero-shot	Puede haber (1) algunos ejemplos de ironía en el siguiente guion. ¿Puedes identificar si hay ejemplos de ironía verbal (2)? Identifícalos todos (3).	Clasifica los ejemplos de ironía que has identificado en el guion en las siguientes categorías: hipérbole, metáfora, exclamación, pregunta retórica y símil. Puede haber más de un ejemplo de ironía en cada categoría.
Few-shot	Puede haber algunos ejemplos de ironía en el siguiente guion. ¿Puedes identificar si existen ejemplos de ironía verbal? Identifícales todos.	Clasifica los ejemplos de ironía que has identificado en el guion en las siguientes categorías: hipérbole, metáfora, exclamación, pregunta retórica y símil. Puede haber más de un ejemplo de ironía en cada categoría. Aquí tienes ejemplos de cada categoría: - Hipérbole: «This suitcase weighs a ton» (The suitcase doesn't actually weigh a ton, but it feels very heavy.) - Metáfora: «He is a lion on the battlefield» (This metaphor compares a person to a lion, suggesting that they are fierce and powerful in a particular situation.) - Exclamación: «What an amazing view!» - Pregunta retórica: «Do you want to lose weight without feeling hungry?» (This question is used to persuade the audience to consider a particular approach to weight loss, rather than expecting a direct answer.) - Símil: «As quick as a flash»

En la fase *zero-shot,* se proporciona un *prompt* al modelo que le guía para completar una tarea específica sin ningún ejemplo, mientras que en el *few-shot* se proporciona al modelo un pequeño número de ejemplos (o solo uno) para una tarea en particular, junto con un *prompt.* En cuanto al primer *prompt,* que es exactamente el mismo tanto en *zero-shot* como en *few-shot,* hay algunas palabras o expresiones que son relevantes para conseguir que el bot de conversación con IA haga lo que se le pide:

1. No se asume que hay ironía verbal. El bot de conversación está forzado a detectarla.
2. Es necesario especificar que tiene que encontrar ironía verbal. De lo contrario, el bot puede identificar otros tipos de ironía, como la ironía situacional.
3. Se pide al bot que los identifique todos, ya que solo podría identificar un número específico de ejemplos (solo cinco o seis).

Del mismo modo, en lo que se refiere al segundo *prompts,* se incluye esta frase: «Puede haber más de un ejemplo de ironía en cada categoría». Si no se añade esto, el bot de conversación puede clasificar solo un ejemplo en cada categoría. Por consiguiente, la primera fase es la identificación o detección de la ironía verbal, mientras que la segunda fase consiste en la clasificación de los ejemplos identificados en las cinco categorías que se han mencionado previamente. Esto aporta información sobre la detección de la ironía y sobre la capacidad del bot para entender y clasificar el lenguaje en dos contextos diferentes: con y sin ejemplos de cada categoría. Así, estos dos pasos dan una visión más amplia de su capacidad con relación al lenguaje.

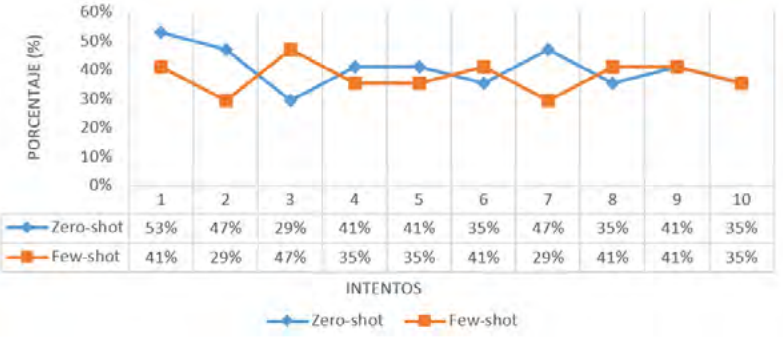
La interacción consiste en diez intentos para cada bot de conversación y cada tipo de *prompt* (*zero-shot* y *few-shot*) en días diferentes, pues los resultados varían. Además, se utiliza un nuevo chat para cada intento, a fin de evitar la influencia entre las interacciones. Todos los resultados se han registrado en una tabla que incluye información sobre el número de ejemplos irónicos identificados en cada bot de conversación y de qué categoría son, el porcentaje de aciertos, el número de errores, el número de categorías correctamente clasificadas junto con el porcentaje de aciertos y el número de errores, la fecha de acceso y el enlace de la conversación. En ocasiones, el número de ejemplos de ironía identificados y el número de ejemplos clasificados correctamente no coinciden, ya que podría haber un ejemplo que se clasificara correctamente, pero que no se identificase en la primera fase. De igual modo, algunas de las categorías pueden superponerse.

3. Resultados

3.1. Resultados de la detección de la ironía por parte de ChatGPT

Como ya se ha mencionado anteriormente, ChatGPT utiliza la tecnología GPT-3.5, que no está tan avanzada como GPT-4. En cuanto al *prompt* de identificación, el bot reconoce, de promedio, el 41 % de los ejemplos de ironía del guion (aproximadamente, 7 de 17) en *zero-shot,* y el 38 % en *few-shot* (aproximadamente, 6 de 17), como se ilustra en el gráfico 1.

Gráfico 1. Fase 1 – Identificación por ChatGPT



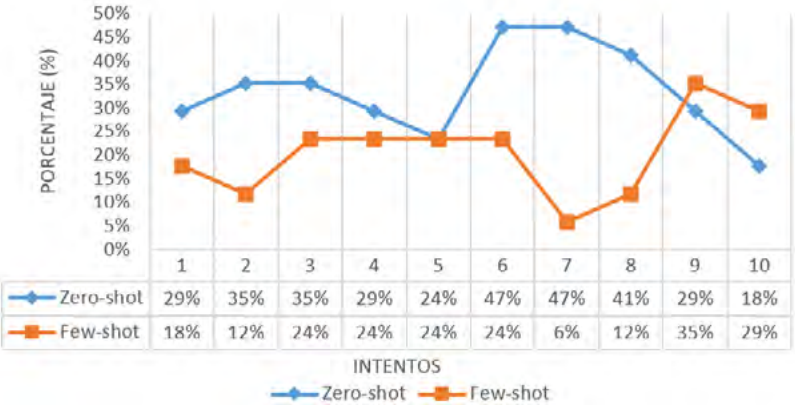
Los resultados son bastante similares y, tras analizar todos los datos, la conclusión es que la diferencia entre *zero-shot* y *few-shot* (*zero-shot* identifica un 3 % más de ejemplos) no es significativa, dado que el número de ejemplos identificados depende exclusivamente del intento, y esta primera tarea es exactamente la misma en ambos tipos de *prompt*. Es decir, el número de expresiones irónicas identificadas en cada intento varía, pero no hay ninguna razón específica para esta variación de resultados. Además, el número de errores de exceso de identificaciones fluctúa de 0 a 3, con un promedio de 1 error en ambos tipos de *prompt*.

En cuanto a los mejores resultados en *zero-shot,* el primer intento fue el más preciso en lo relativo a la detección de ironía. Pese a que tuvo un error de exceso de identificación, ChatGPT fue capaz de identificar 9 enunciados irónicos de 17 (53 %). Por su parte, el peor intento fue el tercero, con solo 5 ejemplos de ironía detectados (29 %). Asimismo, identificó demasiadas expresiones irónicas y cometió 3 errores. Por lo que se refiere a *few-shot,* el mejor intento fue el tercero, con un 47 %

de los ejemplos identificados, y solo tuvo un error de exceso de identificación. Los intentos 2 y 7 detectaron solo el 29 % de las expresiones irónicas, lo cual coincide con el peor porcentaje de *zero-shot*. Además, hubo una actualización en la página web el 16 de mayo de 2024, pero no es relevante en relación con este estudio, dado que los resultados siguieron siendo los esperados a partir de los primeros intentos.

De promedio, en la fase de clasificación, el 34 % de los ejemplos de ironía (anteriormente) identificados se han clasificado correctamente en *zero-shot* (aproximadamente, 6 ejemplos en cada intento), y el 21 % en *few-shot* (aproximadamente, 3-4 ejemplos en cada intento), mostrando mejores resultados en *zero-shot* (gráfico 2).

Gráfico 2. Fase 2 – Clasificación por parte de ChatGPT



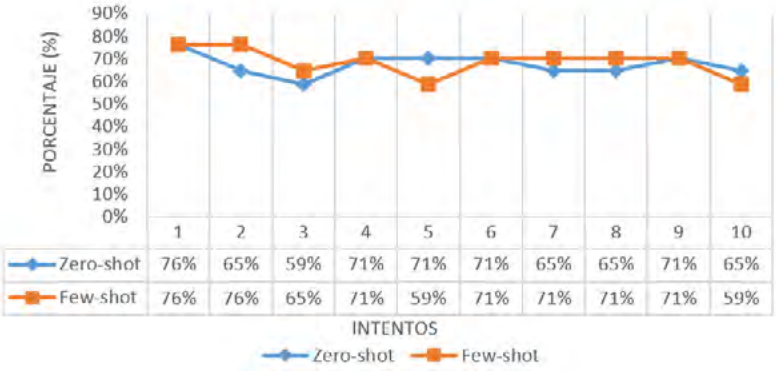
Además, hay, de media, 3 errores de exceso de identificación en *zero-shot* y 2 en *few-shot*. En esta fase, parece que *few-shot* mejora con el tiempo, mientras que *zero-shot* hace lo contrario. En ocasiones, los números no coinciden entre las fases 1 y 2, puesto que podría haber algunas expresiones irónicas clasificadas correctamente, pero que no se identificaron en el paso anterior, como en el intento 6 en *zero-shot* (hay 2 símiles que se clasifican en la segunda fase, pero que no se detectaron en la fase de identificación). Además, dado que algunas categorías pueden superponerse, los números pueden variar de la fase 1 a la fase 2.

En general, ChatGPT y la tecnología GPT-3.5 han identificado menos de la mitad de las expresiones irónicas presentes en el guion. Sin embargo, no tienden a cometer demasiados errores de exceso de identificación, lo cual es positivo. Por otra parte, también han sido capaces de clasificar menos de la mitad de los ejemplos de ironía, y el bot tiende a ser más preciso cuando no se proporcionan ejemplos (es decir, con *zero-shot*). También suele haber más errores que en la primera fase (entre 0 y 6). Normalmente, los errores son clasificaciones erróneas de frases no irónicas que no se han identificado previamente, frases irónicas que no coinciden con esa categoría, o una justificación incorrecta.

3.2. Resultados de la detección de la ironía por parte de Copilot

Microsoft Copilot utiliza la tecnología GPT-4 y la búsqueda en línea. Por tanto, se esperaban mejores resultados en cuanto a la detección de ironía, como se muestra en el gráfico 3.

Gráfico 3. Fase 1 – Identificación por parte de Copilot



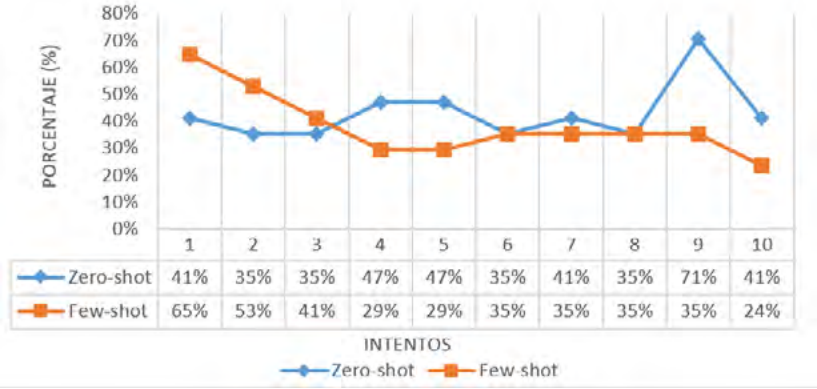
	1	2	3	4	5	6	7	8	9	10
Zero-shot	76%	65%	59%	71%	71%	71%	65%	65%	71%	65%
Few-shot	76%	76%	65%	71%	59%	71%	71%	71%	71%	59%

Copilot ha podido detectar una media del 68 % de los ejemplos de ironía en *zero-shot* (aproximadamente, 11-12 de 17) y del 69 % en *few-shot* (aproximadamente, 12 de 17), mostrando resultados similares en la fase de identificación. Además, el bot de conversación no ha cometido ningún error de exceso de identificación en ninguno de los intentos. El intento 1 en *zero-shot* mostró los mejores resultados en cuanto a la detección de ironía, dado que identificó a 13 enunciados irónicos de 17 (76 %). En cuanto a *few-shot,* los mejores intentos son el primero y el

segundo, ya que también se han detectado 13 ejemplos. Por otra parte, solo se reconocieron 10 expresiones irónicas en el tercer intento de *zero-shot*. En *few-shot,* los intentos 5 y 10 son los menos precisos, ya que Copilot encontró 10 ejemplos en el guion. Con todo, dado que el número fluctúa de 10 a 13, los resultados son bastante estables a lo largo de los intentos.

En cuanto a la segunda parte de la interacción (gráfico 4), el 43 % (de promedio) de las expresiones irónicas se agruparon correctamente en *zero-shot* (aproximadamente, 7 ejemplos en cada intento) y el 38 % en *few-shot* (aproximadamente, 6 ejemplos en cada intento). De nuevo, algunos ejemplos que se clasifican en esta fase pueden no ser identificados en la anterior. Además, los *prompts* con *zero-shot* muestran mejores resultados en lo que se refiere a la clasificación de las expresiones irónicas. Como en el caso de ChatGPT, las categorías que más se han identificado son la metáfora y la exclamación.

Gráfico 4. Fase 2 – Clasificación por parte de Copilot



Sin embargo, de promedio, existe el mismo número de errores de clasificación errónea en los dos tipos de *prompts* (aproximadamente, 3 errores en cada intento), que fluctúa del 1 al 5. En esta fase, aparte de clasificar las categorías, el bot también ofrece una breve definición del recurso lingüístico.

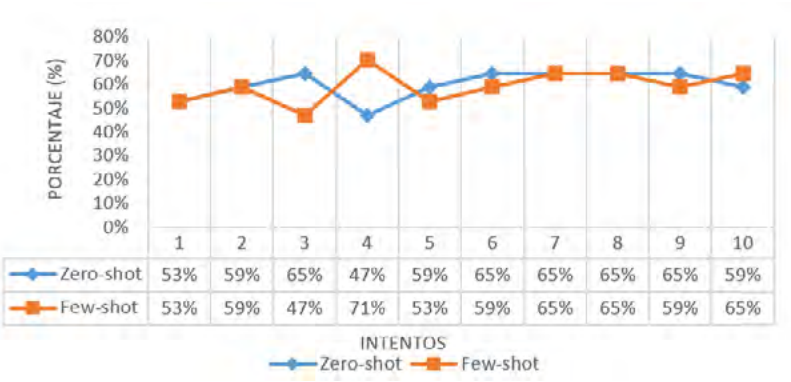
En general, Copilot y la tecnología personalizada GPT-4 son prometedores. Como cabía esperar, ha superado a ChatGPT 3.5 en relación con

la detección de ironía y clasificación del lenguaje. El bot de conversación ha podido identificar entre 10 y 13 ejemplos de ironía, lo que representa el 68-69 % de todas las expresiones irónicas del guion. Además, suele ser bastante más preciso, puesto que no comete errores de exceso de identificación en la primera fase. En cuanto a la segunda fase, el bot ha clasificado casi la mitad de los ejemplos de ironía (anteriormente) identificados. En definitiva, en relación con el objetivo de este estudio, Copilot ha mostrado los mejores resultados.

3.3. Resultados de la detección de la ironía por parte de Gemini

Gemini funciona con tecnología PaLM-2 y es un LLM distinto que no está relacionado con la tecnología GPT. Sin embargo, utiliza un motor similar desarrollado por Google y es el bot de conversación con IA más avanzado de la empresa. De ahí que también se esperasen buenos resultados en cuanto a la detección de ironía, como se muestra en el gráfico 5.

Gráfico 5. Fase 1 – Identificación por parte de Gemini

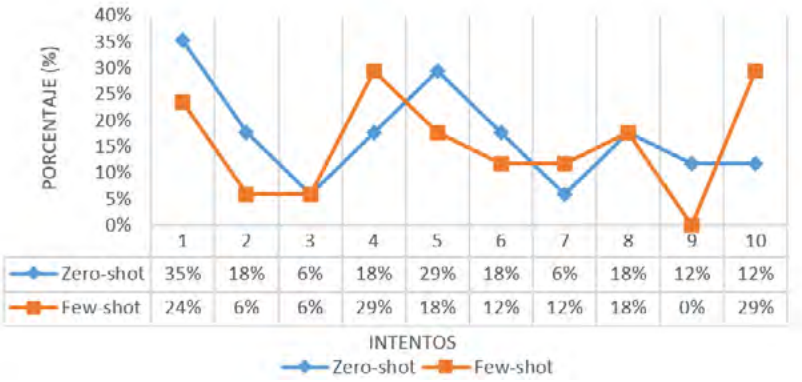


Gemini ha sido capaz de identificar, por término medio, el 60 % de los ejemplos de ironía en *zero-shot* (10 de 17, aproximadamente) y el 59 % en *few-shot* (también 10 de 17 ejemplos en cada intento, aproximadamente). Además, no se han producido errores de exceso de identificación en la mayoría de los intentos. Sin embargo, si un intento incluye algún error, solo hay 1 o 2. Aunque los resultados son bastante estables, el mejor intento es el cuarto en *few-shot,* ya que el bot detectó el 71 % de las expresiones irónicas en el guion. En cuanto a los peores

resultados, se halló el 47 % de los ejemplos de ironía tanto en *zero-shot* como en *few-shot* (intentos 4 y 3, respectivamente). Gemini se actualizó el 14 de mayo, pero este hecho no es relevante en relación con el propósito de esta investigación, puesto que los resultados siguieron siendo los esperados a partir de los primeros intentos.

En cuanto a la clasificación de categorías, Gemini ha clasificado, de media, un 17 % de los ejemplos de ironía (aproximadamente, 3 ejemplos en cada intento) en *zero-shot* y un 15 % en *few-shot* (aproximadamente, 3 en cada intento). En este caso, la diferencia entre ambos tipos de *prompts* no es muy significativa, dado que muestran resultados similares. Además, como se muestra en el gráfico 6, son bastante inestables y tienden a cambiar mucho dependiendo del intento.

Gráfico 6. Fase 2 – Clasificación por parte de Gemini



El número de ejemplos no coincide entre las fases 1 y 2, debido a la superposición de categorías y a la clasificación de expresiones irónicas que no se han identificado previamente. Además, Gemini promedia 1 error de clasificación errónea en ambos tipos de *prompts*. A pesar de ser un bot de conversación con IA avanzado y comparable a GPT-4, ha tenido problemas a la hora de clasificar los ejemplos de ironía en las categorías del estudio. Es decir, pese a haberle especificado que debe clasificar los enunciados irónicos en las cinco categorías, el bot se negó a hacerlo en casi todos los intentos y, finalmente, se vio obligado a hacerlo después de un par de interacciones. Así pues, en lugar de ceñirse

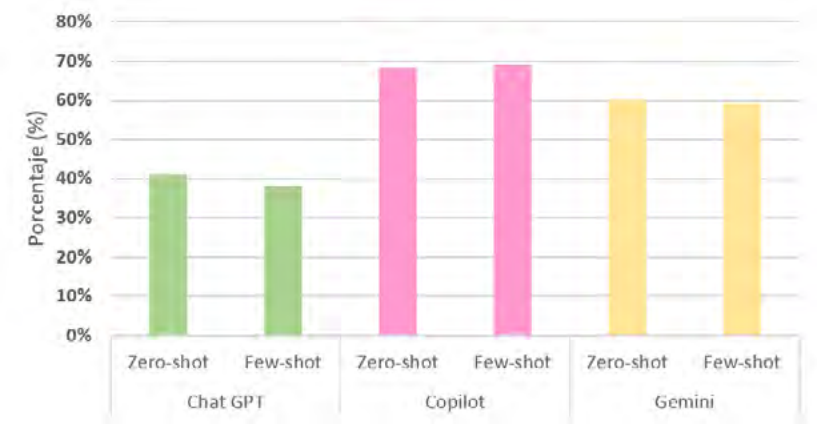
a las instrucciones del usuario, Gemini desafía la clasificación dada y ofrece su propia organización y clasificación de ironía.

En general, Gemini es un bot de conversación con IA interesante, puesto que es capaz de detectar más de la mitad de los ejemplos irónicos presentes en el guion. Sin embargo, también desafía al *prompt* dado para la clasificación de categorías y muestra su propia organización en relación con los enunciados irónicos, que normalmente tiene relación con el sarcasmo, el contexto y las contradicciones.

4. Discusión

Si se comparan todos los bots de conversación con IA, se llega a la conclusión de que Copilot supera a ChatGPT y Gemini (gráfico 7) en la detección de ironía. Además, ChatGPT ocupa la tercera posición y existe una diferencia de un 20 % con Gemini y de un 30 % con Copilot, aproximadamente. Se esperaban mejores resultados de Copilot, pues se trata de una versión personalizada de GPT-4 y también incluye la búsqueda en internet. Además, es la más estable identificando ironía (siempre encuentra entre 10 y 13 ejemplos de los 17 presentados en este estudio).

Gráfico 7. Fase 1 – Detección de ironía

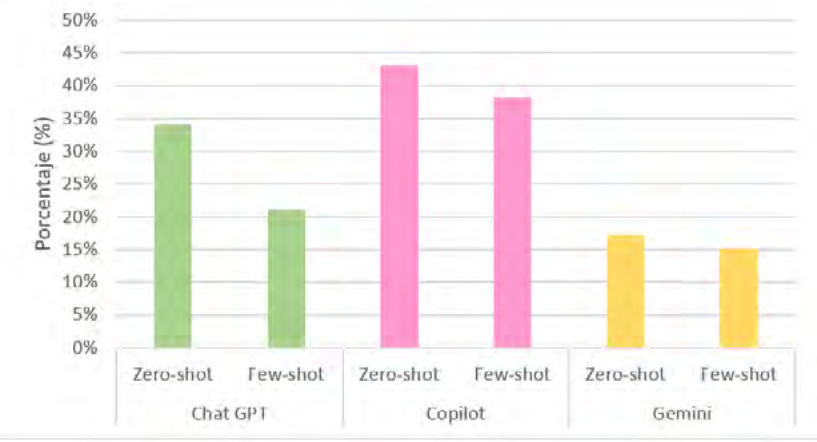


Las metáforas y las exclamaciones son las categorías que más se han detectado en todos los bots de conversación. Además, la variación de resultados entre todos los intentos de cada bot demuestra que pueden

reconocer más enunciados irónicos de los que realmente mencionan. Por ejemplo, si un ejemplo de hipérbole en particular se menciona al menos en uno de los intentos, significa que el bot es capaz de categorizarlo como ejemplo de ironía en general, aunque no se detecte en otros intentos.

En cuanto a la clasificación de las categorías, Copilot también supera a los otros bots. Además, se aprecian diferencias significativas entre los *prompts zero-shot* y *few-shot*. Los datos muestran que la clasificación de categorías funciona mejor en ChatGPT y Copilot cuando no se proporciona ningún ejemplo en el *prompt,* mientras que la diferencia entre los tipos de *prompts* no es tan significativa en Gemini, como se representa en el gráfico 8. Por consiguiente, en general, se obtienen mejores resultados si no se dan ejemplos y los bots se basan solo en los datos en los que se han entrenado.

Gráfico 8. Fase 2 – Clasificación



En definitiva, Copilot es el bot que mejores resultados ha proporcionado en relación con la detección y la clasificación de ironía, y ChatGPT ha obtenido los peores resultados en la detección de ironía en comparación con los otros bots de conversación con IA, aunque ha superado a Gemini en la clasificación. Dado que GPT-3.5 es el predecesor de GPT-4, se puede observar una cierta evolución y mejora en el modelo. ChatGPT trata con datos almacenados hasta 2021, mientras que Copilot se basa en GPT-4 y otros datos de la búsqueda en internet (Bing).

En otras palabras, existe una mejora de un LLM a otro en relación con la detección de la ironía, ya que Copilot está entrenado con más datos y también se basa en información actualizada del buscador en internet. Por otra parte, Gemini, el bot más avanzado de Google, también es interesante, debido a su tecnología y a un LLM diferente (PaLM-2). Por consiguiente, por más que no supere a Copilot, sus resultados son bastante precisos y mejores que los de ChatGPT. En particular, es importante destacar que este bot va más allá de las instrucciones dadas a través de un *prompt* del usuario. Es decir, mientras que ChatGPT y Copilot se adhieren a las instrucciones, Gemini es capaz de desafiar las instrucciones del usuario si piensa que el *prompt* es incorrecto, dado que se niega a clasificar los ejemplos de ironía en las categorías dadas y ofrece su propia clasificación para identificar expresiones irónicas.

5. Conclusión y limitaciones del estudio

Esta investigación ha explorado el potencial de la IA para detectar la ironía a través de algunos de los bots de conversación con IA más comunes entre los usuarios (OpenAI ChatGPT, Microsoft Copilot y Google Gemini), puesto que la ironía es un recurso lingüístico muy utilizado en el lenguaje humano para transmitir ideas complejas. Por ello, es interesante que los bots estén entrenados para poder captar expresiones irónicas y tener una mejor comprensión de los *prompts* dados por el usuario. Después de utilizar el guion preparado para poner a prueba las IA, se ha encontrado que Copilot supera a los otros bots en relación con este propósito, y también es mejor clasificando expresiones irónicas en comparación con los demás. Además, en lo que se refiere a la clasificación en las cinco categorías, el *prompt zero-shot* funciona mejor que el de *few-shot* en ChatGPT y Copilot, mientras que la diferencia no es tan significativa en Gemini. Dicho de otro modo, los bots tienen mejores resultados si no se proporcionan ejemplos de las categorías en el *prompt*. Además, parece que la razón por la que Copilot es mejor que ChatGPT en la detección de ironía es que GPT-3.5 es el predecesor de GPT-4. La tecnología GPT-3.5 está entrenada solo con datos disponibles hasta 2021, mientras que la versión personalizada de GPT-4 de Copilot se ha entrenado con mayores cantidades de datos y también se basa en la información disponible en la búsqueda de internet en Bing. En cuanto a

Gemini, que es el bot más avanzado de Google y que se basa en PaLM-2, también tiene buenos resultados con relación a la identificación de la ironía. Sin embargo, y en comparación con los demás, este bot desafía las instrucciones del usuario. Es decir, Gemini se niega a clasificar los ejemplos en las categorías especificadas. Después de analizar todos los datos, puede concluirse que los resultados son positivos y prometedores. Si bien es cierto que los bots presentan algunas limitaciones (errores de exceso de identificación, ejemplos que no se identifican, etc.), su potencial en relación con la ironía, el lenguaje no literal y la pragmática está mejorando, puesto que GPT-4 y PaLM 2, que están entrenados con más datos, superan a GPT-3.5, que es más antiguo. En otras palabras, los bots parecen cada vez más precisos y están entendiendo mejor el lenguaje natural, incluyendo enunciados irónicos. Como consecuencia, esto puede implicar que futuros LLM y bots de conversación con IA sean aún superiores.

Este estudio ha contribuido a los campos de la lingüística, la pragmática y las tecnologías del lenguaje. Además, podría ser de interés para el análisis de sentimientos y la investigación sobre análisis de opiniones. Como los LLM siguen evolucionando y aparecen nuevos bots, es difícil mantenerse al día y hay poca investigación sobre la detección de la ironía y los bots de conversación con IA. Este estudio pretende llenar un nuevo vacío creado a partir de la explosión de la IA, el PLN y los LLM: cómo abordan estas tecnologías el lenguaje complejo (como, por ejemplo, las expresiones no literales). Sin embargo, la metodología utilizada tiene algunas limitaciones. Se ha creado un guion con una variedad de ejemplos de ironía, pero se implementan artificialmente en el texto y no representan una conversación natural o el *prompt* real de un usuario. Además, el potencial de los bots de conversación se restringe a este guion y a los *prompts* utilizados específicamente para este propósito. Por lo que hace a futuros trabajos de investigación en esta área, sería interesante analizar una muestra más representativa de ejemplos de ironía utilizados en un abanico más amplio de contextos. Dado que el número de ejemplos de ironía en el guion es bastante bajo, la cantidad de expresiones irónicas debería aumentar. De la misma forma, un tipo diferente de *prompt* puede alterar los resultados. Asimismo, deberían explorarse más a fondo otros bots, así como seguir explorando los de este estudio a fin de analizar cambios y tendencias, y contribuir a la evolución continua de los bots, los LLM y las tecnologías del lenguaje.

5. CONCLUSIONES

Los tres trabajos presentados en este volumen han analizado el uso de la IA y, en concreto, de los LLM, para la retroacción autocorrectiva en textos y para la detección de la ironía.

En el caso de los trabajos sobre la retroacción autocorrectiva, cabe destacar la diferencia que existe en la retroacción proporcionada dependiendo de la lengua de los textos. A pesar de haber analizado el propio LLM (ChatGPT 3.5), en el caso del inglés las correcciones realizadas son, en general, acertadas en todos los niveles explorados. Sin embargo, en el caso de las redacciones en catalán, muchas de las correcciones son erróneas, poco precisas o incluso se proporcionan respuestas en español. Aquí, pues, se ve claramente el papel primordial que tienen los datos con los que los sistemas han sido entrenados, no solo en términos de calidad, sino también, especialmente, de cantidad. El número de datos disponibles en inglés supera ampliamente el de los datos disponibles en catalán, lo cual tiene una clara incidencia en los resultados obtenidos. También cabe destacar que, en el caso del inglés, los sistemas analizados son capaces de dar retroacción positiva al usuario, mientras que en el caso del catalán esto no es así. De nuevo, los datos con los que se han entrenado estos sistemas para ambas lenguas son, posiblemente, responsables de este tipo de respuesta. El número de textos y de materiales disponibles para la enseñanza del inglés como L2 o como lengua extranjera es ampliamente superior al número de materiales para la enseñanza del catalán. Así, el sistema está más familiarizado con la tarea de dar retroacción en inglés que en catalán.

A partir de estos resultados, pueden sugerirse algunas líneas de mejora para los LLM aplicados a lenguas con menos recursos, como el catalán. Por ejemplo, entrenar estos modelos con corpus normativos específicos en catalán, incluyendo textos y materiales docentes, podría mejorar su precisión. Asimismo, adaptar las respuestas del sistema a distintos niveles educativos (por ejemplo, usuarios de primaria, secundaria o universidad) permitiría ofrecer una retroacción más ajustada al contexto. También sería útil desarrollar módulos que incorporaran estrategias de retroacción positiva y constructiva, similar a como se hace para el

inglés, para fomentar la motivación y la mejora progresiva en lenguas minorizadas.

En el plano pedagógico, parece que estos sistemas y, en especial, en el caso de lenguas más minoritarias como el catalán, conviene utilizarlos para una corrección inicial, pero sigue siendo necesaria la supervisión e intervención humanas para garantizar una corrección precisa y contextualmente adecuada. Esto indica que es preciso formar al alumnado y al profesorado en el uso de estas herramientas, así como evaluar cómo integrarlas en la enseñanza.

En cuanto a la detección de la ironía, el estudio que se presenta en este volumen demuestra que los sistemas de IA todavía han de recorrer mucho camino hasta que puedan detectarla. Detectar la ironía tiene que ver claramente con el conocimiento del mundo y las referencias contextuales y culturales del usuario. Según los resultados obtenidos en el estudio realizado, la IA todavía no posee dicho conocimiento. Sin embargo, cabe destacar que las versiones más avanzadas (Copilot) llevan a cabo mejor esta tarea que sus predecesores (ChatGPT 3.5), lo que vuelve a indicar que la disponibilidad de grandes cantidades de datos con los que entrenar a los sistemas resulta fundamental. La detección de la ironía es un aspecto clave que los sistemas de IA han de mejorar para tener éxito en la interacción entre el ser humano y la máquina.

Como propuesta de mejora, sería necesario crear corpus específicamente anotados con ejemplos de ironía y otros fenómenos pragmáticos (por ejemplo, presuposiciones, implicaturas, eufemismos, hipérboles), incorporando también metadatos culturales y contextuales. Además, un entrenamiento híbrido que combine datos lingüísticos con información sobre conocimiento del mundo podría ayudar a los modelos a captar mejor la discrepancia entre el significado literal y la intención comunicativa. Asimismo, la integración de herramientas multimodales (por ejemplo, con soporte visual) podría enriquecer la interpretación de este tipo de contenido.

Las aportaciones recogidas en este volumen muestran cómo el uso de la IA en el análisis lingüístico puede ayudar a definir criterios claros para entender y abordar cuestiones complejas del lenguaje, a la vez que ofrece ejemplos concretos de aplicación didáctica en contextos universitarios. La exploración de esta temática en los trabajos de fin de grado

de los estudios de Filología abre nuevas vías de reflexión sobre la enseñanza y la evaluación de la producción escrita, e invita a repensar el papel de las herramientas digitales en la formación de los futuros profesionales de la lengua. Se trata de un campo que empieza a tomar forma y que, a partir de la experiencia acumulada en prácticas docentes, puede convertirse en un terreno fértil para la innovación y el pensamiento crítico.

REFERENCIAS

Almelhes, S. A. (2023). Review of Artificial Intelligence Adoption in Second-Language Learning. *Theory and Practice in Language Studies*, *13*(5), 1259-1269. https://doi.org/10.17507/tpls.1305.21

Banasik, N. (2013). Non-literal speech comprehension in preschool children– An example from a study on verbal irony. *Psychology of Language and Communication*, *17*(3), 309-324.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Amodei, D. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33. https://doi.org/10.48550/arXiv.2005.14165

Byron, C., Wallace, D. K. C., Kertz, L. y Charniak, E. (2014). Humans Require Context to Infer Ironic Intent (so Computers Probably do, too). En: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 512-516. Baltimore, Maryland. Association for Computational Linguistics.

Chen, J., He, X. y Miyao, Y. (2024). Language Model Based Unsupervised Dependency Parsing with Conditional Mutual Information and Grammatical Constraints. En: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6355-6366. Mexico City, México. Association for Computational Linguistics.

Ettinger, A., Hwang, J. D., Pyatkin, V., Bhagavatula, C. y Choi, Y. (2023). «You Are An Expert Linguistic Annotator»: Limits of LLMs as Analyzers of Abstract Meaning Representation. En: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8250-8263. Singapur. Association for Computational Linguistics. 10.18653/v1/2023.findings-emnlp.553

Frenda, S. (2016). Computational rule-based model for irony detection in italian tweets. En: *EVALITA 2016*, 1749, pp. 1-6.

González, J. A., Hurtado, Ll. F. y Plà, F. (2019). Elirf-upv at irosva: Transformer encoders for spanish irony detection. En: *Proceeding of the Iberian Language Evaluation Forum (IberLEF 2019)*. https://ceur-ws.org/Vol-2421/IroSvA_paper_4.pdf

Google (2024). Gemini [Large Language Model]. https://gemini.google.com

Guo, K. y Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 1-29. https://doi.org/10.1007/s10639-023-12146-0

Hromei, C. D., Croce, D. y Basili, R. (2025). U-DepPLLaMA: Universal Dependency Parsing via Auto-regressive Large Language Models. *IJCoL*, 10-11. https://doi.org/10.4000/125nm

Ilic, S. Marrese-Taylor, E., Balazs, J. y Matsuo, Y. (2018). Deep contextualized word representations for detecting sarcasm and irony. En: *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 2-7. Bruselas. Association for Computational Linguistics.

Kristiawan, D., Y., Bashar, K. y Pradana, D. A. (2024). Artificial intelligence in English language learning: A systematic review of AI tools, applications, and pedagogical outcomes. *The Art of Teaching English as a Foreign Language* (TATEFL), *5*(2), 207-218. https://doi.org/10.36663/tatefl.v5i2.912

Machado, M. y Ruiz, E. (2024). Evaluating large language models for the tasks of PoS tagging within the Universal Dependency framework. En: *Proceedings of the 16th International Conference on Computational Processing of Portuguese - vol. 1*, pp. 454-460. Santiago de Compostela. Association for Computational Linguistics.

Microsoft (2023). *Copilot* [Large Language Model]. https://copilot.microsoft.com

Morreel, S., Verhoeven, V. y Mathysen, D. (2024). Microsoft Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. *PLOS Digital Health*, *3*(2), e0000349.

Navarro-Carrascosa, C. (coord.) (2024). La enseñanza del Español como Lengua Extranjera y la Inteligencia Artificial: perspectivas y retos. *Doblele: revista de lengua y literatura*, *10*.

OpenAI (2023). *ChatGPT* [Large Language Model]. https://chat.openai.com

Pérez-Núñez, A. (2024). ChatGPT in Spanish language instruction: exploring AI-driven task generation and its implications for teaching practices. *Journal of Spanish Language Teaching, 11*(1), 61-82. https://doi.org/10.1080/23247797.2024.2366053

Potamias, R. A., Siolas, G. y Georgios Stafylopatis, A. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309- 17320.

Reyes, A., Rosso, P. y Veale, T. A. (2013). multidimensional approach for detecting irony in Twitter. *Lang Resources & Evaluation, 47*, 239-268. https://doi.org/10.1007/s10579-012-9196-x

Schmidt, T. y Strasser, T. (2022). Artificial intelligence in foreign language learning and teaching: a CALL for intelligent practice. *Anglistik: International Journal of English Studies, 33*(1), 165-184. https://angl.winter-verlag.de/data/article/11071/pdf/92201014.pdf

Tang, Y. y Chen, H. (2014). Chinese irony corpus construction and ironic structure analysis. En: *Proceedings of COLING 2014, 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1269-1278.

Taskıran, A. y Goksel, N. (2022). Automated feedback and teacher feedback: Writing achievement in learning English as a foreign language at a distance. *Turkish Online Journal of Distance Education*, *23*(2), 120-139.

Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics*, *4*(1). https://doi.org/10.1016/j.acorp.2024.100089

Van Hee, C., Lefever, E. y Hoste, V. (2018a). We usually don't like going to the dentist: Using common sense to detect irony on Twitter. *Computational Linguistics*, *44*(4), 793-832.

Van Hee, C., Lefever, E. y Hoste, V. (2018b). Semeval-2018 task 3: Irony detection in English tweets. En: *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 39-50.

Wei, P., Wang, X. y Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: a randomized controlled trial. *Front. Psychol.*, *14*, 1249991. DOI: 10.3389/fpsyg.2023.1249991

Xiang, R., Gao, X., Long, Y., Li, A., Chersoni, E., Lu, Q. y Huang, C. R. (2020). Ciron: a new benchmark dataset for Chinese irony detection. En: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5714-5720. Marsella. European Language Resources Association.

Yu, D., Li, L., Su, H. y Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis. The case of apology. *International Journal of Corpus Linguistics*, *29*(4), 534-561. https://doi.org/10.1075/ijcl.23087.yu

NORMAS PARA LOS COLABORADORES

https://bit.ly/3DKFESh

EXTENSIÓN

Las propuestas de cada cuaderno no podrán exceder **la extensión de 50 páginas (en Word)**, salvo excepciones, unos 105.000 caracteres; espacios, referencias, cuadros, gráficas y notas incluidos.

PRESENTACIÓN DE ORIGINALES

Los textos han de mostrar, en formato electrónico, un **resumen** de unas diez líneas y tres palabras clave, no incluidas en el título. Igualmente, han de contener el **título**, un ***abstract*** y tres ***keywords*** en inglés.

Respecto a la **manera de citar y a las referencias bibliográficas**, se han de remitir a las utilizadas en este cuaderno.

EVALUACIÓN

La aceptación de originales se rige por el **sistema de evaluación externa por pares**.

Los originales son leídos, en primer lugar, por el **Consejo de Redacción**, que valora la adecuación del texto a las líneas y objetivos de los cuadernos y si cumple los requisitos formales y el contenido científico exigido.

Los originales se someten, en segundo lugar, a la **evaluación de dos expertos** del ámbito disciplinar correspondiente, especialistas en la temática del original. Los autores reciben los comentarios y sugerencias de los evaluadores y la valoración final con las correcciones y cambios oportunos que se han de aplicar antes de ser aceptada su publicación.

Si los cambios exigidos son significativos o afectan a buena parte del texto, el nuevo original se somete a la evaluación de dos expertos externos y de un miembro del Consejo de Redacción. El proceso se lleva a cabo como «doble ciego».

REVISORES

https://bit.ly/3oF4izw

LA INTERSECCIÓN ENTRE LA IA Y EL ANÁLISIS LINGÜÍSTICO. EXPERIENCIAS DE EVALUACIÓN DE LA IA EN LA DESCRIPCIÓN Y EL ANÁLISIS LINGÜÍSTICOS DEL INGLÉS Y EL CATALÁN

NATALIA JUDITH LASO MARTÍN
ELISABET COMELLES PUJADAS
(COORDINACIÓN)

Octaedro Editorial

Institut de Desenvolupament Professional
UNIVERSITAT DE BARCELONA