ChatGPT en la creación de exámenes de alemán como lengua extranjera

ChatGPT in creating exams on German as a foreign language

Celia Llabrés Llovet
Universitat de les Illes Balears
celia.llabres3@estudiant.uib.cat
https://orcid.org/0009-0002-8692-8135

Recibido: 09/07/2024
Aceptado: 10/11/2024
DOI: https://dx.doi.org/10.12795/mAGAzin.2024.i32.03

Resumen:
ChatGPT se disfraza de ser humano. Por este motivo, se pretende comprobar si esta interfaz es capaz de crear pruebas de evaluación de alemán como lengua extranjera de calidad. El principal objetivo de esta investigación es evidenciar si ChatGPT tiene la capacidad de sustituir la labor docente de crear exámenes.
La metodología del presente estudio ha consistido en el envío de dos encuestas similares, una a profesorado y otra a alumnado de alemán del ámbito universitario, para comprobar si son capaces de distinguir entre textos producidos por humanos, específicamente exámenes de la Escuela Oficial de Idiomas, o textos similares creados por ChatGPT.
En ocasiones, ChatGPT comete errores y expresa incoherencias, lo cual no puede permitirse en la evaluación de conocimientos de lenguas extranjeras, sobre todo, si los humanos son capaces de crear esos mismos exámenes de manera más eficiente y rápida.

Palabras clave: ChatGPT, EOI, coherencia, profesorado, alumnado

Abstract:
ChatGPT pretends to be a person. Therefore, the aim of this article is to prove if the chatbot can create quality exams on German as a foreign language. The main purpose of the investigation is to prove if ChatGPT is able to substitute the professors' task of creating exams and, consequently, substitute them in this specific duty.
The methodology conducted has consisted of two similar surveys sent to German professors and university students, one to each, in order to observe if they are capable of distinguishing between texts written by humans, specifically exams from the Spanish Official School of Languages, or similar texts written by ChatGPT.
Aside from the personal views of the participants, ChatGPT both makes mistakes and is occasionally incoherent, which cannot be permitted in the assessment of foreign language knowledge, mainly if humans can create these same exams in a faster and more efficient manner.

Keywords: ChatGPT, Spanish Official School of Languages, coherence, professors, students

El 30 de noviembre de 2022 los humanos nos despertamos con una nueva realidad, con una idea de la que pocos habían oído hablar y que, desde entonces, no para de repetirse en todos los medios e instituciones: ChatGPT había llegado para quedarse, el mayor exponente de la inteligencia artificial (IA) actual.

1. El concepto de ChatGPT

El éxito de ChatGPT, un *Generative Pre-trained Transformer* desarrollado por la empresa OpenAI, fue inminente, superando en pocos meses las cifras que las plataformas que lideraban los récords del momento, TikTok e Instagram, habían tardado años en conseguir (Ichbiah 2023: s.p.). La diferencia notoria es el formato conversacional de esta interfaz inteligente, pues da la impresión de que se trate de una persona experta en cualquier tema debido a la coherencia y a la cohesión de la mayoría de sus respuestas, aunque hay que tener en cuenta que realmente no entiende el mensaje que comunica (*ibid*), sino que se nutre de muchísimas bases de datos, con las cuales intenta proporcionar la respuesta que aparenta ser más natural en el idioma que se le pide, no necesariamente correcta.

La gran novedad es que ChatGPT «can generate new content rather than simply analyze existing data» (Baker 2023: s.p.), lo cual se equipara a lo que, hasta el momento, se entendía que solo podía hacer un humano: crear nuevo contenido «desde cero». Su mayor objetivo es la cohesión, mientras que «en muchos temas es poco precisa» (Fernández 2024: s.p.).

1.1. La aplicación de ChatGPT en la educación

En particular, el campo de la educación convive con esta nueva realidad: ChatGPT se utiliza a diario en las aulas con el fin de facilitar el trabajo. Por ello, es urgente el estudio de sus beneficios, inconvenientes y limitaciones; para poder, así, incluirlo en el aula de manera beneficiosa a todos los participantes del proceso educativo.

Al analizar el rendimiento de otras herramientas que se habían diseñado «para llevar a cabo tareas pedagógicas» (Sabzalieva & Valentini 2023: 29), hasta ahora siempre se resaltaba el papel fundamental que, a cualquier nivel en cualquier ámbito educativo, tiene el profesorado. En cuanto al estudiante, para conseguir información valiosa por parte de ChatGPT, Baker (2023: s.p.) menciona que debe preguntar a la interfaz la respuesta correcta, lo cual implica un análisis crítico previo por parte del alumnado que desee utilizar dicha herramienta como ayuda en sus labores. De hecho, Engelke & Engelke (2023: s.p.) enfatizan la importancia del uso de estas tecnologías en clase, ya que argumentan que los alumnos que no se vean impulsados a utilizarlas se verán rezagados frente aquellos que sí aprovechen los beneficios de su uso porque «a major AI service like ChatGPT has the potential to be a great educational tool if utilized well» (Kim 2023: s.p.).

¿Pero qué ocurrirá con los profesores: dejarán de hacer (parte de) su trabajo? ¿Se reducirá plantilla en las instituciones educativas? De ahí surge una creciente preocupación. Ahora bien, la respuesta a esta incógnita tan solo nos la podrá dar el tiempo; así y todo, estudios como el de Loos *et al.* (2023) concluyen que ChatGPT no es capaz de reemplazar al maestro: «The answers given by ChatGPT seemed correct, but they were also superficial» (2023: 13). La mayoría de participantes de otras investigaciones como la de Alenizi *et al.* (2023: 17) recomiendan no reemplazar la enseñanza personal y el apoyo individualizado, sino que la IA sirva de suplemento.

No se debe olvidar que ChatGPT comete errores ocasionalmente, lo cual podría evitarse con la preeminencia de un profesor. Asimismo, «al menos por ahora, ChatGPT no puede sustituir a la creatividad humana y [a]l pensamiento crítico» (Sabzalieva & Valentini 2023: 14); no es capaz —por lo menos, de momento— de proporcionar «feedback», guiar o animar individualmente a los alumnos, además de carecer de las habilidades que debería tener un profesor como la empatía o la inevitable interacción humana (CE Noticias Financieras 2023: 1-3). Mondal *et al.* (2023: 204) describen el proceso de aprendizaje como una posible experiencia emocional, en la cual los estudiantes podrían necesitar un profesor que les apoye y oriente. De lo que parece no caber duda es que «the role of humans will change as technology advances» (Kim 2023: s.p.), aunque Baker (2023) es firme en la convicción de que «educators are the ones who move a society forward and enable it to adapt» (s.p.).

Otras investigaciones (Sabzalieva & Valentini 2023: 13) sugieren evaluaciones presenciales o

modificaciones del tipo o del formato de preguntas para aprovechar ChatGPT en las instituciones educativas, en vez de prohibirlo tajantemente. Incluso, Ray (2023: s.p.) recoge diferentes prácticas relacionadas con la docencia que puede llevar a cabo ChatGPT con seguridad, como la enseñanza personalizada mediante materiales y actividades y el apoyo al docente a través de recomendaciones del plan de estudios.

En resumen, los estudios mencionados se centran en las capacidades que ChatGPT puede tener a la hora de enseñar; sin embargo, no tienen en cuenta aspectos que también juegan un papel fundamental en esta profesión: la evaluación de la materia que se ha enseñado previamente en el aula.

2. El presente trabajo

Ante estas innovaciones recientes se hace necesario investigar de qué manera el profesorado puede aprovechar la IA en la evaluación del alumnado. En este trabajo en concreto, se ha elegido analizar la puesta en práctica de ChatGPT por dos razones en especial: por su popularidad extendida a prácticamente todas las naciones y a todos los grupos de edad y sociales, y por ser la evidente demostración de hasta dónde puede llegar la IA (Ichbiah 2023: s.p.). Es, indudablemente, la aplicación de la IA de la que más han hablado los medios, probablemente, por la incertidumbre que genera y, consecuentemente, por el pánico que puede llegar a causar.

Asimismo, se ha trabajado con la versión gratuita de ChatGPT pues, de esta manera, está al alcance de cualquiera, independientemente del nivel socioeconómico del individuo. Hay que tener en cuenta que la versión de pago de ChatGPT tiene un mejor motor (se trata de GPT-4, en vez del modelo GPT-3.5 de la versión gratuita), además de un «acceso prioritario a nuevas funciones» (Fernández 2023: s.p.).

Además, este trabajo se centra en la enseñanza de lenguas extranjeras, ya que el desempeño del trabajo de profesores de lenguas extranjeras afecta prácticamente a todas las personas, independientemente de su campo de especialización en el mundo laboral, pues todos aprendemos, hemos aprendido o aprenderemos algún idioma diferente a nuestra lengua materna. En este caso, el posible uso de ChatGPT para evaluar las competencias aprendidas en el aula también afecta al alumnado. Específicamente, la lengua extranjera con la que se va a investigar el desempeño de ChatGPT en el mundo profesional del docente es el alemán, al tratarse de un idioma cada vez más hablado y, por ende, estudiado; y que, además, en cierta medida, se aleja del constante foco de estudio de lenguas extranjeras que se centra en el inglés.

Así, mediante el presente trabajo, se pretende analizar los puntos fuertes y débiles de los exámenes de nivel B1 de alemán que crea ChatGPT con los mismos enunciados que un examen modelo de una Escuela Oficial de Idiomas de España mediante variables cuantitativas y cualitativas por parte de profesores que imparten esta lengua y alumnos o exalumnos que han recibido clases de la misma.

La hipótesis previa al estudio es positiva para las TICS, a la vez que negativa para el sector profesional: la autora espera que ChatGPT sea capaz de crear exámenes a partir de instrucciones claras y específicas. Esto podría suponer una reducción del trabajo de los docentes en lo que, hasta ahora, formaba parte de su jornada laboral. No obstante, se parte de la premisa de que habría que revisar las propuestas de ChatGPT, es decir, podría caber la posibilidad de que hubiera que realizar pequeñas modificaciones —teniendo en cuenta que la corrección de los exámenes no debería suponer un mayor consumo de tiempo que el de confeccionarlas—.

El método seguido por este estudio es, por orden cronológico, el envío de un cuestionario a profesores y otro similar a alumnos de lengua alemana en niveles educativos superiores. En segundo lugar, se indicarán y analizarán las respuestas recibidas. Por último, se concluirá con los principales hallazgos y con las limitaciones que se han dado en esta investigación y sus circunstancias en concreto.

3. Metodología

3.1. Los cuestionarios

Con la finalidad de comparar los resultados que proporciona ChatGPT con los que ha creado un humano, se pidió a esta interfaz la realización de ejercicios similares a los que aparecieron en el examen de la Escuela Oficial de Idiomas de La Rioja en 2022 del nivel B1 de alemán (de acuerdo con el Marco Común Europeo de Referencia para las lenguas MCER), disponible en su sitio web (véase enlace en el «Anexo I»).

Se eligió este determinado nivel (y no uno superior) porque una mayor dificultad alargaría el tiempo

necesario para rellenarlo debidamente. En segundo lugar, tres estudiantes que han cumplimentado la encuesta se encuentran actualmente cursando clases de alemán B1+; de esta manera, se ampliaba el número de participantes del cuestionario. Por último y más importante, ChatGPT se negó, en repetidas ocasiones, a proveer textos de una longitud considerable, propios de niveles superiores a B1. De hecho, como se comentará más adelante, se le tuvo que solicitar que aumentara la extensión de los textos porque, de lo contrario, los participantes de los cuestionarios hubiesen sabido, sin necesidad de reflexión, qué texto había creado ChatGPT y qué otro, la EOI.

Adicionalmente, los exámenes para alcanzar los certificados de los niveles a partir del B1 «están regulad[o]s y organizad[o]s por las administraciones educativas de acuerdo a unos principios comunes» (EDUCAGOB s.f.). En consecuencia, dicho nivel se podría considerar el mínimo para poder investigar, ya que cumple ciertos requisitos establecidos por las autoridades españolas.

Además, se seleccionó esta determinada Escuela Oficial española de manera aleatoria. De los cinco participantes que han trabajado en una Escuela Oficial de Idiomas en territorio español, dos de ellos se encontraban en Castilla y León, uno en Córdoba, otro en Madrid y un último participante en las Islas Baleares, Cataluña y La Rioja. Únicamente uno de ellos ejercía en el momento en el que se realizó el estudio.

La comunicación con ChatGPT fue en español en todo momento. Para cada una de las partes que se solicitaba se iniciaba un nuevo chat para que no hubiese interferencias con el resto de resultados, es decir, se evitó en todo momento que el hilo de la conversación que se estaba manteniendo con ChatGPT pudiese influir en la creación de las nuevas pruebas. En cuanto al procedimiento, se proveía a ChatGPT del enunciado exacto de que constaba cada una de las partes del examen original; seguidamente, se solicitaba que crease una actividad de la misma índole con ese mismo enunciado. En la mayoría de casos, era necesario solicitarle en repetidas ocasiones que alargara el texto o añadiera huecos en blanco o preguntas para igualarlo al ejercicio de la EOI original y que, consecuentemente, la diferencia entre ambos fuera mínima.

Ambos cuestionarios, de Formularios de Google, recogían las mismas secciones con las mismas preguntas comparando los exámenes de la EOI y de ChatGPT, en las cuales se debía contestar cuál se creía que había creado quién, junto a su debida justificación. Sin embargo, se dirigió, por una parte, un cuestionario a profesores y, por otra parte, otro a alumnos, en el que la diferencia radicaba en las preguntas demográficas, que se explicarán más adelante, en esta misma sección de Metodología.

> Ambos cuestionarios, uno dirigido a profesores y otro a alumnos, comparaban el examen de la EOI y el de ChatGPT.

Aparte de las preguntas personales, cada una de las partes del cuestionario eran las diferentes secciones del examen: la comprensión de textos escritos, la prueba de mediación escrita en alemán y en español, la producción y coproducción de textos escritos y la producción y coproducción de textos orales. Igualmente, se han descartado la comprensión de textos orales y la prueba de mediación oral, ya que ChatGPT no es capaz de crear contenido multimedia, es decir, audios e infografías respectivamente, las cuales aparecían en el examen seleccionado para la investigación. A la hora de solicitarle dichas secciones, su respuesta fue la siguiente: «Lo siento, pero como modelo de texto, no puedo proporcionar archivos de audio. Sin embargo, puedo ofrecerte transcripciones […]» (OpenAI 2023, 31 de diciembre).

Las respuestas cualitativas se han formateado de manera libre, es decir, los participantes podían extenderse en sus justificaciones cuanto quisieran y no estaban obligados a contestar en ningún caso; en otras palabras, las preguntas se contestaban sin un mínimo ni máximo de caracteres de manera voluntaria, ya que se pretendía que cada participante enfocase la respuesta a su manera y sin ninguna idea preconcebida. No obstante, algunos participantes se han limitado a contestar si creían que ChatGPT o la EOI había creado el examen sin explicar el porqué en algunas de las preguntas, lo cual supone una limitación para este trabajo. La investigación tan solo ha tenido

en cuenta las respuestas de calidad, es decir, las que incluían justificaciones que reflejaban de manera clara una reflexión.

Es pertinente aclarar que las instrucciones y las preguntas de ambos cuestionarios se escribieron tanto en español como en inglés, ya que se anticipó la participación de docentes y alumnos que no fuesen exclusivamente de nacionalidad española. Paralelamente, se debe mencionar que no todos los participantes han contestado en el apartado de mediación escrita en español dado que son extranjeros y no entienden el idioma español o sus conocimientos son limitados e insuficientes para responder adecuadamente. En este caso, tan solo se han tenido en cuenta las respuestas de los participantes que entendieron la tarea y la han respondido.

Finalmente, se optó por recabar los datos de la investigación por medio de cuestionarios, ya que se trata de una metodología asincrónica, es decir, los participantes pueden responder donde y cuando puedan, además de tomarse el tiempo que consideren necesario. Se debe tener en cuenta que la mayoría de los participantes se encontraban en diferentes lugares de España a la hora de rellenar el formulario o, incluso, otros se hallaban en Alemania o Canadá, por mencionar dos países a modo de ejemplo. Por lo tanto, la flexibilidad fue una prioridad a la hora de optar por este método de investigación. Asimismo, una encuesta aseguraba que constara por escrito el anonimato de los participantes y la única y exclusiva finalidad de los datos proporcionados para el objeto de investigación.

En el «Anexo II» se adjunta el cuestionario enviado al profesorado, mientras que el «Anexo III» está dedicado al cuestionario enviado al alumnado, con sus respectivos enlaces.

3.1.1. Cuestionario dirigido al profesorado

En primer lugar, es preciso puntualizar que este cuestionario en concreto estuvo disponible a través de un enlace durante enero y febrero de 2024, y que respondieron dieciséis mujeres y cuatro hombres. En un primer momento, se contactó con 82 profesores universitarios en total, mayoritariamente de universidades públicas españolas, aunque también de universidades alemanas y canadienses y de algunas privadas españolas, de forma aleatoria. En un segundo momento, se envió el enlace a las dos responsables de las Áreas de Filología Alemana y de Traducción e Interpretación del Departamento de Filología y Traducción de la Universidad Pablo de Olavide (UPO), con la petición de que hiciesen llegar el enlace a ambas áreas en su conjunto. Cuando esto ocurrió, una de las profesoras, perteneciente a la Asociación de Germanistas de Andalucía (AGA), trasladó la petición de que se rellenase la encuesta a los miembros de dicha asociación. En cuanto al medio de contacto por el cual se envió dicho cuestionario, se trató del correo electrónico. En la mayoría de los casos, no se obtuvo respuesta ninguna por parte de los docentes. Incluso, hubo repuestas de profesores que no se consideraban aptos para dicha investigación por diferentes motivaciones.

Todos los encuestados son o han sido profesores de alemán que ejercen o han ejercido su profesión en la universidad, mayoritariamente en el ámbito público (a excepción de tres, que tan solo han puntualizado ejercer en secundaria y bachillerato). Sin embargo, este no es o ha sido necesariamente su único trabajo. En cuanto al país en el que han trabajado, en total, veinticuatro de los participantes alguna vez han desempeñado un empleo en territorio español, aunque cinco han ejercido únicamente en España. A este quinteto se le une un participante, sumando seis en total, que no ha trabajado nunca en un país de habla alemana.

En la sección de preguntas personales, también se les preguntó por la incorporación de la IA en sus clases. Prácticamente la mitad de los participantes, doce, nunca la utilizaron por falta de interés, de necesidad, de información y/o de tiempo o de fiabilidad. Una de las participantes contestó que apenas la había utilizado. Once docentes habían utilizado la herramienta a veces y solo dos la utilizaban con frecuencia en o para sus clases. Los usos que más se repiten entre las respuestas son la traducción automática y la creación de materiales, como ejercicios de gramática o de redacción. De entre todos los participantes, catorce comprenden el rango de edad de 50 a 65 años; mientras que once, de 31 a 50. Tan solo uno es menor de 30.

3.1.2. Cuestionario dirigido al alumnado

Los participantes totales de este cuestionario fueron doce; la mitad, hombres y la otra mitad, mujeres. El rango de edad comprendía de los 20 a los 26 años. Cinco de los encuestados llevaban más de cinco años aprendiendo el idioma alemán y otros cinco, entre tres y cinco años; mientras que dos, entre uno y dos años. Estos dos últimos, sorprendentemente, no

estaban incluidos en los tres que consideraban tener un nivel de B1 en la lengua. También tres participantes consideraban tener un nivel B2; y dos, un C2. Finalmente, cuatro de ellos valoraron su conocimiento en un C1.

Con relación a los estudios universitarios que cursaban, ocho participantes estudiaban Grados que incorporan el alemán, como Traducción e Interpretación, mientras que cuatro estudiaban Medicina, aunque habiendo cursado clases de alemán en la universidad en varias ocasiones.

El enlace de acceso al cuestionario se distribuyó a través de la aplicación WhatsApp y los mensajes directos de Instagram; estuvo disponible en enero y febrero de 2024.

4. Discusión de los resultados

Con la finalidad de analizar los resultados obtenidos en los cuestionarios, se expondrán en gráficos los porcentajes de respuestas correctas e incorrectas de los participantes (además de «NS/NC» (No sabe/ No contesta), lo cual corresponde a todo tipo de ambigüedades) a la hora de reconocer qué exámenes estaban creados por la EOI y qué otros, por ChatGPT, teniendo en cuenta el total de respuestas, es decir, conjuntamente tanto de la encuesta de profesorado como también la de alumnado. Cada figura corresponderá a una sección del examen diferente. Tras exponer los datos cuantitativos, se analizarán las justificaciones que se han dado al respecto.

4.1. Comprensión de textos escritos

En todas las partes de «Comprensión de textos escritos», más de la mitad de los participantes ha identificado correctamente si se trataba de una máquina o de un humano (véase Figuras 1, 2 y 3). Además, esta sección contiene un alto número de respuestas y un bajo número de «NS/NC». Probablemente esto se deba a que es la sección inicial. Igualmente, es el apartado con mayor variedad de respuestas, ya que en secciones posteriores algunos participantes han repetido las mismas justificaciones que indicaron en esta sección.

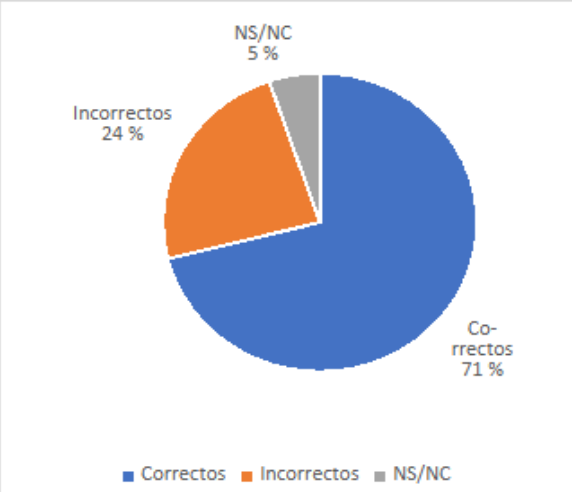

Figura 1. Comprensión de textos escritos 1

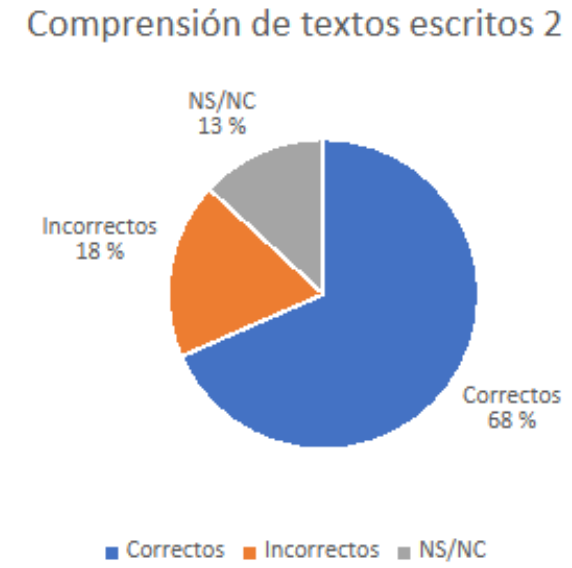

Figura 2. Comprensión de textos escritos 2

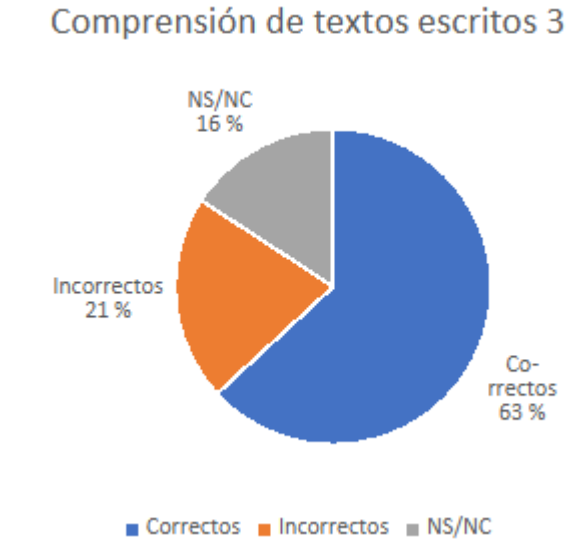

Figura 3. Comprensión de textos escritos 3

Llabrés Llovet, Celia: *ChatGPT en la creación de exámenes de alemán como lengua extranjera*
Magazin 32, ISSN 1136-677X. 2024/25, pp. 24-38. https://dx.doi.org/10.12795/mAGAzin.2024.i32.03

4.1.1. Respuestas correctas

La mayoría de los participantes que han elegido la opción correcta se han justificado por los errores que comete ChatGPT (ortográficos, gramaticales, de vocabulario e, incluso, incoherencias), sobre todo en la primera tarea de la comprensión escrita, pero también encontramos ejemplos en la parte 3, en la que conviene destacar que las soluciones a los huecos estaban escritas en el propio texto, a continuación del propio hueco.

Se debe mencionar que algunos de los que han contestado (erróneamente) que dicho texto era de la EOI se han dado cuenta de estos errores también; sin embargo, han seguido optando por la opción incorrecta, probablemente, por otros matices que han considerado más desventajosos y que se comentarán más adelante.

Precisamente, relacionado con dichos errores, uno de los encuestados que ha elegido la opción correcta ha puntualizado que una desventaja de utilizar esta interfaz es que «el proceso de revisión, a veces, resulta más largo que el de búsqueda de textos ya terminados». Paradójicamente a lo comentado en la sección de «Introducción» del presente trabajo, un considerable número de participantes encuentran que el texto de ChatGPT resulta «carente de cohesión», no solo en esta sección, sino que se repite en la mayoría de ellas; algunos añaden que esto es debido a la falta de conectores, lo cual atenta contra uno de los principales objetivos de ChatGPT: la coherencia a la que Ichbiah (2023) aludía, con la que nos hace pensar que se trata de un humano.

Otras justificaciones de entre los que distinguieron adecuadamente los autores de los textos son que la prueba de ChatGPT resulta estereotipada a causa de «aspectos muy concretos fáciles de localizar en el texto» o una «típica división en párrafos, cada uno con un título» sin «progresión temática o argumental». Al contrario de la EOI, que desarrolla una historia, la IA «constantemente […] menciona las mismas ideas».

Por añadidura, se acusa en diversas ocasiones a la IA de desarrollar el tema de forma muy general. De igual forma, se ha calificado como «demasiado plano», «superficial y trivial», «monótono, con el mismo tipo de frases», de «formulación artificial», «demasiado objetivo y descriptivo» y de «lengua […] neutra». Una de las razones por las que ChatGPT genera este tipo de textos es, según dos participantes, la falta de «adornos literarios que sí están presentes en el otro texto». Además, la EOI ha incluido, según otro par de encuestados, «elementos culturales clave», que ChatGPT ha omitido. Otras calificaciones del texto son simplicidad e, incluso, en una ocasión, brevedad.

En el plano léxico, se ha definido como «muy descriptivo y narrativo». Hay quienes consideran que no siempre se ha elegido adecuadamente. Además, un encuestado puntualizaba que una desventaja de los exámenes que crea la IA es que «you don't have full control of the vocabulary included».

Incluso, quienes ya habían observado el *modus operandi* de ChatGPT en un pasado han remarcado que no suele incluir discurso directo, referencias reales ni utilizar «comillas para proponer expresiones pronunciadas por otras personas». Varios participantes han seguido la línea de uno de ellos, que comentaba: «No me parece posible que un *chatbot* sea capaz de escribir tan convincentemente tomando el punto de vista de un personaje público y conocido». De hecho, a ChatGPT le delata su impersonalidad, «generalidades y sin ninguna referencia clara a personas o lugares concretos», tan solo «un cúmulo de lugares comunes». Otro usuario añadía que ChatGPT suele presentar la información «a modo de lista o mediante 'bullet points' […] teniendo la desventaja de que suele fallar en originalidad».

Estos mismos participantes elogiaban los exámenes de la EOI poniendo en valor virtudes, de las cuales consideran que ChatGPT, en consecuencia, carece. Las respuestas, en este caso, eran más unánimes. La que más se repetía eran la mayor dedicación al texto, lo cual se reflejaba en el hecho de que el examinando se vea obligado a invertir un mayor esfuerzo en la comprensión porque «debe tener capacidad de interpretación», de «relacionar unas partes con otras»; de la misma manera que «no está tan focalizado en la comprensión de una sola palabra concreta» y «puede reflejar mejor el grado de comprensión que pueda haber tenido el estudiante o la estudiante», con lo cual «evalúa el contenido». En cambio, la IA escaseaba de «razonamiento lógico, una deducción por contexto o ir un poco más allá de la literalidad» porque «las preguntas se han tomado del texto al pie de la letra y no hay ninguna 'trampa'». Más de uno añadía que el orden las respuestas de ChatGPT tenía una «conexión muy directa entre el comienzo de cada párrafo y la respuesta que le corresponde». Igualmente, en varias ocasiones se repitió la diferencia de nivel entre el texto y las preguntas de ChatGPT, es decir, la mayor

complejidad del texto comparado con la simplicidad de las preguntas.

Por añadidura, se destacaba con frecuencia la naturalidad del texto de la EOI: la cercanía «al habla del día a día», «a la realidad y a la persona», con un «registro muy coloquial y cultural» y la espontaneidad. Otros adjetivos que también describen este segundo punto han sido «*lebendig*» (dinámico) y «humano». Mientras que ChatGPT mostraba una «sencilla y repetitiva estructura y contenido» o un «lenguaje tipo máquina». Incluso, un encuestado que no acertó de quién era el texto afirmaba que el texto de la EOI (que pensaba que era de ChatGPT) «busca gramática correcta según manuales y no tanto con instinto nativo». Otro participante mencionaba que precisamente los «giros lingüísticos» eran los que delataban a la máquina.

Otro punto que delataba a la EOI era el ejercicio 0 que se suele escribir al principio a modo de ejemplo para que el estudiante se sitúe y entienda qué se le está pidiendo exactamente, que ChatGPT no incluyó en su examen.

Por último, entre las respuestas correctas, los temas de los textos de la EOI se consideraban de actualidad, lo cual no sucedía en el caso de la IA.

De entre los que sí reconocieron la autoría de los textos, uno de los participantes acusaba a ChatGPT de sintético, mientras que otro elogiaba a la EOI por eso mismo.

4.1.2. Respuestas erróneas

En cambio, los participantes que se equivocaron dudaban en muchos casos y sus argumentos para defender su elección no eran tan sólidos como los de sus compañeros. No obstante, en una respuesta (errónea), se utilizaba «*clearly*» para indicar la convicción de la respuesta (incorrecta); por lo tanto, no en todos los casos se dudó de la elección, aun cuando esta fue incorrecta.

Respecto al léxico, este grupo de participantes creyeron que el examen de la IA era de la EOI por el nivel de vocabulario descrito como variado, «más homogéneo y bastante cercano a los potenciales estudiantes» o «*advanced*». A diferencia del grupo de respuestas correctas, la relativa sencillez se ha considerado un rasgo positivo. Quienes lo han justificado comentaban que «en ellas [las preguntas] los alumnos no se ven obligados a construir frases completas, sino que solamente deben realizar una tarea de elección múltiple entre tres posibles respuestas».

Asimismo, de entre estas respuestas, se resaltaba positivamente la estructura, la claridad, la fluidez y, por extraño que parezca, la cohesión del texto de ChatGPT. De la misma manera, se valoraba positivamente el uso de conectores «habitualmente empleados en los cursos de alemán» y «*a variety of advanced grammatical structures*». Por lo que se refiere a la generalización y a la división por apartados con su respectivo título, que las respuestas correctas tanto reprochaban, han sido objeto de elogio en este conjunto de contestaciones.

Otro punto que un encuestado ha matizado es el uso por parte de ChatGPT de «expresiones típicas que se buscan reconocer en ambientes didácticos», es decir, el uso de conectores compuestos o de verbos junto a preposiciones, que se usan en ciertos niveles y están, por lo tanto, aplicados en el examen.

De la EOI, teniendo en cuenta que creían que se trataba de ChatGPT, han criticado el uso de anglicismos y abuso de genitivos, pero, sobre todo, una mayor complejidad y longitud del texto. En último lugar, cinco participantes acusaban a la EOI de falta de un hilo conductor. Uno incluso se dejó engañar y pensaba que «*it sounds like if it were a translation of an English text*».

4.2. Mediación escrita en alemán

En este apartado, las respuestas han sido más breves, comparadas con las de la «Comprensión de textos escritos». A ello se suma que la naturaleza de esta actividad es diferente: de participación más activa por parte del alumnado que realiza la prueba y, por lo tanto, mucho más corta. A pesar de ello, el número de contestaciones correctas continúa siendo más de la mitad.

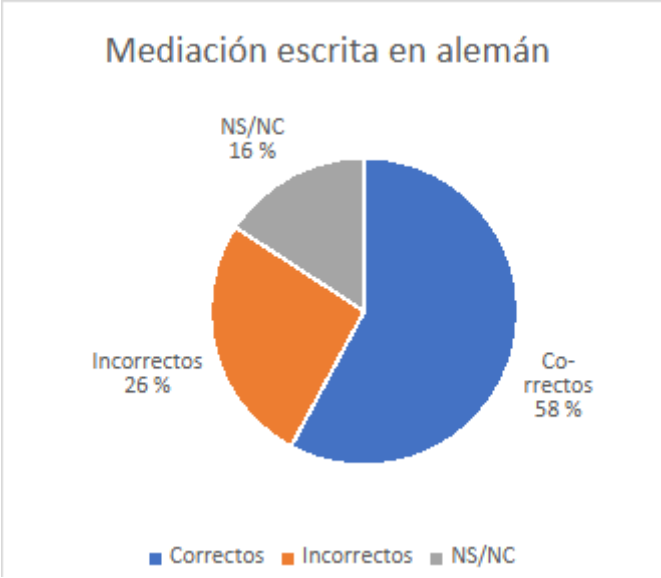

Figura 4. Mediación escrita en alemán

4.2.1. Respuestas correctas

Entre las respuestas correctas, se comentaba que ChatGPT era de «frases demasiado largas» y, aun siendo un fragmento breve, dos participantes han identificado errores en el texto que produjo la IA.

Los puntos nuevos que aportan cuatro encuestados que han identificado con éxito la autoría de los textos son el hecho de que no se trata de un ejercicio de mediación verdaderamente: «No se cumple el objetivo de la mediación (aclarar/explicar/resumir una idea de un texto en un idioma a otro). Aquí, ChatGPT pide una redacción». Por tanto, la máquina carece, claramente, de adecuación en el caso de la mediación en alemán. Por el contrario, el texto de la EOI «permite al alumno expresar ideas originales y mostrar conocimientos en un área más especializada», lo cual «busca una implicación más personal [u] original».

La mayor ventaja que parecía mostrar el examen de la EOI era una anotación al final en mayúsculas: «Schreiben Sie ihr eine E-Mail (100-120 Wörter), in der Sie ihr KURZ DAS WICHTIGSTE erklären». Seis participantes han destacado este rasgo como el mayor indicador de que «se percibe una mano humana reparando o previniendo problemas que ya ha vivido en otras ocasiones previas». A diferencia de las demás respuestas correctas, dos participantes (que, además, habían contestado correctamente en todas las partes de la sección anterior) admitían que, esta vez, les parecía más difícil que en la «Comprensión de textos escritos» identificar a ChatGPT. Sin embargo, ha sido la ya mencionada anotación en mayúsculas al final la razón por la que uno de ellos se ha decantado por la opción correcta.

En relación a esto, se añadía que la información de la EOI es relevante y clara para realizar el ejercicio, además de haber «más matices, más instrucciones», a la par que se expresa en «frases más comprimidas», mientras que un participante se atrevió a calificar de «infantil» el texto de ChatGPT.

4.2.2. Respuestas erróneas

De entre las respuestas incorrectas, por un lado, cuatro han considerado que las mayúsculas a las que se acaba de hacer mención «quitan seriedad y profesionalidad al texto» y no parecen adecuadas en este contexto. Además, consideraban que el «*kurz*» agravaba la informalidad: «¿por qué '*kurz*' si ya se indica el número de palabras?». Como resultado, en su conjunto, «las instrucciones finales son indeterminadas y redundantes»; además, opinaban que podía generar confusión: «*the exercise isn't formulated very clearly*».

En cuanto al léxico, «*while both Tipps and Tips may be correct, no person writing would include both spellings in a single sentence*». Adicionalmente, dos participantes lo consideraban erróneo: «Uso incoherente de la palabra *Tipps* (correcto es con dos p)», afirmación que puede interpretarse como que hay quienes relacionan la incoherencia con la autoría de ChatGPT —debe puntualizarse que la palabra con una sola "p" es obsoleta—. Adicionalmente, las respuestas erróneas de la encuesta han considerado que ChatGPT destacaba por claridad y concisión: «students […] are told exactly what to do». Asimismo, encuentran que el tema es *alltagsbezogen* (sobre el día a día), que «empatiza con el estudiante y traslada la tarea a su mundo cotidiano (ej.: la fiesta de cumpleaños de una amiga)».

Finalmente, se ha criticado la elección del tema (puestos de trabajo y el mundo laboral), al tratarse de una «situación [que] no es tan habitual para los estudiantes que a menudo van a las EEOOII; muchos de ellos son todavía adolescentes y están por lo tanto alejados del mundo laboral […]». Sin embargo, este argumento se desvaloriza cuando se lee el siguiente apartado, cuyos exámenes, tanto el de la EOI como el de la IA, tratan precisamente la misma temática.

4.3. Mediación escrita en español

En esta sección, la cifra de respuestas correctas disminuye a poco menos de la mitad, mientras que la de

erróneas aumenta a 14, siendo esta la sección de mayor número de incorrecciones. Muchos participantes admiten haber dudado de esta parte.

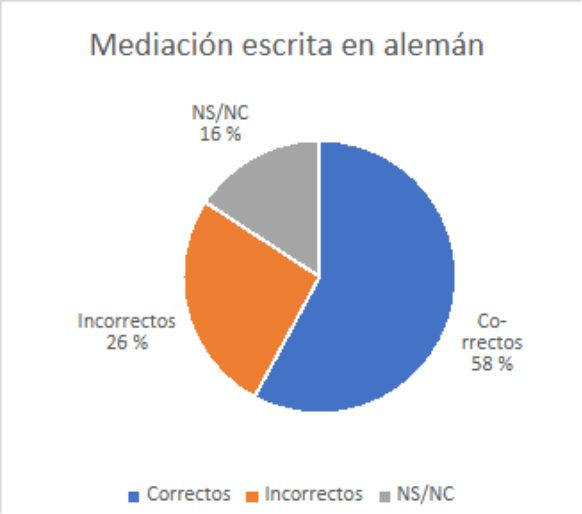

Figura 5. Mediación escrita en español

4.3.1. Respuestas correctas

En esta instancia, ChatGPT escogió un tema de la misma índole que el que se le presentó de la EOI, por lo tanto, el propio tema no fue un factor decisivo para los encuestados. Sin embargo, se le ha dado importancia a cómo se ha presentado el contenido: por parte de las respuestas correctas, se ha observado que ChatGPT había organizado mejor el texto, con su introducción y su conclusión, pero de la misma manera que lo hace siempre, «esperable» y «cliché», lo cual delata su autoría.

Además, se vuelve a repetir la falta de naturalidad por su parte («frases un poco frías», «sencillas, cortas, poco o nada subordinadas. Frases que un humano escribiría unidas, aquí aparecen constantemente separadas por un punto»); especialmente, debido a la formalidad en la que se presenta la información, mientras se resaltan las expresiones idiomáticas por parte de la EOI.

Es precisamente esta formalidad la que le resta el aspecto humano a ChatGPT: «registro más elevado y poco adecuado a la actividad que se pedía». A diferencia de la EOI, «se concentra únicamente en el aspecto profesional y no en el emocional». Como resultado, se incide, de nuevo, en la mayor complejidad, ya que «al alumno le puede resultar complicado sintetizar las ideas de este texto».

Por último, el hecho de que el ejercicio de la EOI es más adecuado para una actividad de mediación vuelve a aparecer en este caso.

4.3.2. Respuestas erróneas

Por lo que se refiere a las respuestas erróneas, la mayoría han considerado que ChatGPT ha apostado por una más organizada y clara estructura de la información, la cual se ha considerado «más completa y eficiente para sacar chicha para la mediación». Aparte, los títulos resaltados en negrita han llevado a confundirles y creer que se trataba de la EOI cuando no era así (como sí ha pasado en secciones previas).

4.4. Producción escrita

En esta penúltima parte, 19 de los 27 participantes que contestaron a ambas partes de la «Producción escrita» han contestado de manera correcta a los dos apartados, mientras que el otro 30 % tan solo contestó de manera correcta una de las partes y erróneamente la otra, lo cual evidencia que no todos los participantes estaban convencidos de sus contestaciones.

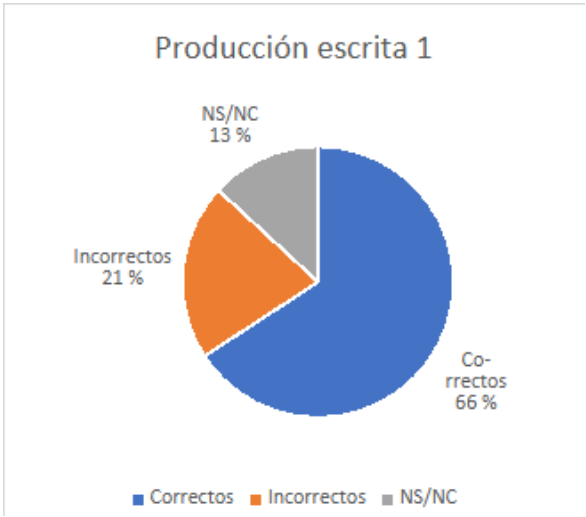

Figura 6. Producción escrita 1

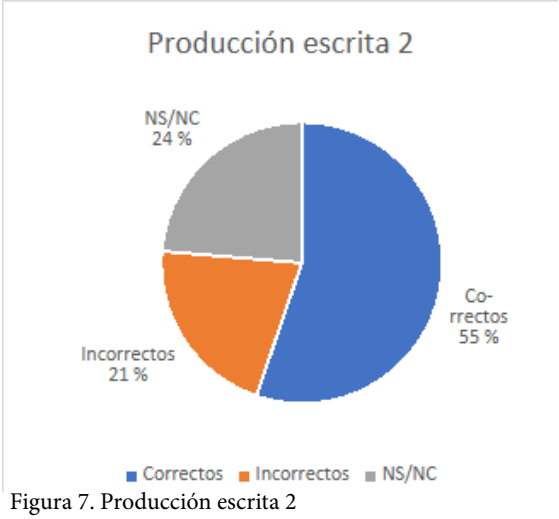

Figura 7. Producción escrita 2

Llabrés Llovet, Celia: *ChatGPT en la creación de exámenes de alemán como lengua extranjera*
Magazin 32, ISSN 1136-677X. 2024/25, pp. 24-38. https://dx.doi.org/10.12795/mAGAzin.2024.i32.03

4.4.1. Respuestas correctas

Las novedades de esta sección respecto a las anteriores son el hecho de que las instrucciones de ChatGPT incluyen que hay que escribir en alemán («Schreibe eine E-Mail auf Deutsch») y comentan lo siguiente: «Achte darauf, einen formellen Ton zu verwenden und deine E-Mail klar und respektvoll zu strukturieren» (Recuerda utilizar un tono formal y estructurar tu email de forma clara y respetuosa), «ya que normalmente presuponen que sabes si es formal/informal»; precisamente, si el estudiante sabe reconocer la formalidad en la que debe escribir es un aspecto que se tiene en cuenta a la hora de evaluar.

Adicionalmente, los participantes con respuestas correctas consideran que es complicado cubrir todos los aspectos que pide el enunciado en el reducido número de palabras que se requieren, mientras que de la EOI comentan que «*the three points listed are a great guidance for the examinees*».

Asimismo, tres participantes han observado que ChatGPT tuteaba al lector, lo cual la EOI no hacía. También se ha comentado el uso —«poco típic[o] en este tipo de ejercicios»— del imperativo. A su vez, se ha calificado en varias ocasiones como «*too formal*» por otros participantes. Se debe remarcar que en este apartado de la encuesta se repite en diversas ocasiones, de nuevo, la estandarización o algún error (en la «Producción escrita 2» el error se sitúa en la primera palabra, por lo que es mucho más visible), mientras que el enunciado de la EOI tan solo se repiten características ya mencionadas, como la «Relevanz im Alltag» (que está relacionado con el día a día) o la especificidad.

En la segunda parte de la producción se ha criticado por parte de prácticamente todos los encuestados que «el enunciado puede ser algo enrevesado». Incluso, un participante ha asegurado que puede llegar a parecer un ejercicio de comprensión lectora. Además, se agrava el problema debido a la mala organización: «solo te ponen dos párrafos con toda la información». En cambio, el texto de la EOI es «corto y claro».

En último lugar, los temas que trata la EOI son los «típicos que se preparan en clase: *Gründe, Meinung, Alternativen, persönliche Situation*...», en comparación con ChatGPT, que «solicita un texto de opinión muy clásico sobre medioambiente. Las opciones de ChatGPT es como si hubiesen estado realizadas a partir de libros de texto de idiomas antiguos».

4.4.2. Respuestas erróneas

En una línea totalmente diferente, las respuestas erróneas parecían estar muy seguras de que los ejercicios producidos con IA habían sido creados por un humano por la supuesta especificidad, complejidad, claridad, estructura y por el «*rich vocabulary*». Es verdad que, según lo comentado en la «Introducción», aparece en más de una ocasión, una vez más, el adjetivo «coherente», referido a ChatGPT, aunque vinculado a la idea de que esa cualidad es más propia de un humano.

Lo que confundió a muchos participantes fue la aclaración al final de las instrucciones, tanto de la «Producción escrita 1» (la que se mencionó anteriormente en el presente trabajo sobre qué tono utilizar a la hora de escribir, además de recordar que se debe estructurar) como la de la 2 («que incluye indicaciones por parte de los examinadores acerca de lo que esperan del ejercicio»). Ambas parecen «da[r] la clave para determinar que hay una mente humana detrás» porque «da a entender que conoce en parte la forma de trabajar de los chicos a los que va dirigida la actividad y trata de reconducir sus posibles errores en la respuesta».

4.5. Producción oral

La «Producción oral» incluye monólogo y diálogo entre los examinandos en un mismo ejercicio, ya que así figuraba en el examen de la EOI. Se debe destacar que esta sección es la que tiene menos respuestas correctas, probablemente, porque tiene el mayor número de «NS/NC», como se puede observar en la Figura 8.

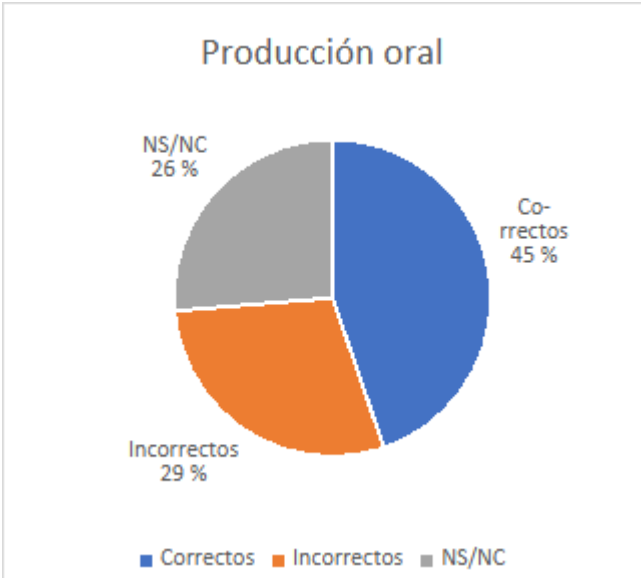

Figura 8. Producción oral

Además, al tratarse del último ejercicio, se repiten todos los motivos con los cuales los participantes ya se han justificado en las secciones anteriores.

4.5.1. Respuestas correctas

La mayoría de las respuestas correctas se justifican por la estructura que suele seguir la EOI, con «instrucciones mucho más claras y concretas», a diferencia de la generalización por parte de ChatGPT que se lleva comentando a lo largo del presente apartado de «Discusión de Resultados». En este caso se observa que el enunciado «se centra solo en la temática de la que tendrán que hablar». Otro punto importante es el hecho de que ChatGPT «se ha olvidado de ustedear». De todas maneras, sigue siendo la EOI la que «busca una implicación más personal [..] [y] la interacción» al «preguntarte por tu opinión» y ChatGPT el que incluye «contenido estándar, esperable».

Por último, un encuestado, aun eligiendo a ChatGPT (correctamente), hace el siguiente comentario: «Buena tarea, bien formulada, me gusta», con lo que se puede inferir que no necesariamente las calificaciones negativas siempre están asociadas a la máquina, sino que los humanos también tenemos la percepción de que existe la posibilidad de que la IA realice la tarea de forma exitosa.

4.5.2. Respuestas erróneas

La confusión en este ejercicio se ha dado, sobre todo, porque el enunciado de ChatGPT estaba en alemán en su totalidad y este hecho se ha asociado a una característica propia de las pruebas de las EEOOII. Contrariamente a las respuestas correctas, las incorrectas encuentran que ChatGPT es claro, corto y preciso; incluso opinan que tiene una «elección de palabras algo más complicada» y que el enunciado del ejercicio «está mejor estructurado».

En última instancia, aun siendo la sección con menos justificaciones novedosas, dos participantes han comentado reflexiones de gran valor. Se debe puntualizar que ambos participantes respondieron de manera incorrecta a la «Producción oral». Uno de ellos aseguraba: «Si el segundo es de la EOI, de verdad tienen que mejorar sus modelos», lo cual reafirma la teoría de que los rasgos negativos se asocian (o quieren asociarse) a la máquina y que, tristemente, como comentaba otro participante, «confío demasiado en la dedicación y habilidad de los examinadores humanos […] e infravaloro la capacidad de ChatGPT de imitar a los humanos que actúan con habilidad y dedicación», que es precisamente la razón principal por la que muchos han respondido de manera errónea a diferentes partes de la encuesta.

5. Limitaciones

Una de las limitaciones de la investigación es la familiaridad que tenían algunos participantes con el tipo de pruebas que se suelen realizar las EEOOII, ya que, en algunos casos, las justificaciones se han limitado a observar el patrón que se suele seguir en vez de analizar el contenido *per se*.

La segunda limitación ha sido el número de participantes en la encuesta, 38 en total (26 profesores y 12 alumnos), lo cual puede no resultar totalmente representativa, teniendo en cuenta la población total de profesorado y alumnado universitario que imparte o aprende alemán como lengua extranjera. Igualmente, entre los participantes, había quienes no hablaban español (o no lo suficiente) y, por ende, no podían responder a la «Mediación en español», lo cual ha aumentado el número de «NS/NC» en dicho apartado. Por añadidura, al tratarse de respuestas abiertas en la encuesta, se han dado ambigüedades (y contradicciones) en algunas contestaciones, es decir, a veces no se entendía con claridad si los encuestados se referían a la EOI o a ChatGPT. Algunos, incluso, no respondían en ocasiones.

Además, este estudio se ha centrado únicamente en la lengua alemana de nivel B1, lo cual no significa necesariamente que sea extrapolable a otros idiomas extranjeros ni a otros niveles. Por ello, la primera propuesta de investigación de cara al futuro es la aplicación de ChatGPT en diferentes lenguas extranjeras y/o que el nivel sea superior, para, además, observar el funcionamiento y la dinámica de la IA según el idioma en el que se le pidan pruebas de evaluación de cierto nivel, sin descartar la posibilidad de que el estudio sea una comparativa entre diferentes lenguas, en la cual se observen las diferencias según el idioma del examen que se le solicite a ChatGPT.

En cuanto al marco teórico del primer apartado, se ha considerado que la fecha de publicación de la mayoría de la bibliografía no podía exceder la fecha en la que surgió ChatGPT (noviembre de 2022). Esta decisión ha supuesto un mayor esfuerzo en la búsqueda de bibliografía anterior a la presente investigación.

Asimismo, este estudio podría aplicarse en un aula en un futuro, con tal de analizar su puesta en práctica en situaciones reales, en las que el alumnado y el profesorado aprendan mutuamente a saber controlar la máquina para, así, evitar que les controle la máquina a los usuarios que intentan beneficiarse de ella.

6. Conclusiones

En primer lugar, llama la atención como unos pocos participantes «elogian» la tarea de ChatGPT, es decir, precisamente porque les parece más correcto que la otra opción, consideran erróneamente que se trata de la IA, lo cual rompe el esquema que han seguido muchos otros, donde los rasgos positivos de los textos se asocian preferentemente con la producción humana.

Entre las tendencias que se han observado, sobre todo en el alumnado, contestaban en la mayoría de los casos correctamente o, al contrario, en la mayoría de los casos de manera errónea. Tan solo tres participantes, un profesor y dos alumnos, han respondido a todo de forma correcta, de entre el total de participantes que han respondido a todas las secciones sin excepción. Eso sí, ningún encuestado ha respondido incorrectamente a todos los apartados.

Por otra parte, se ha observado que, en general, las respuestas del alumnado son, en su mayoría, correctas, mientras que existe una mayor variedad de respuestas —correctas e incorrectas— entre el profesorado, lo cual podría indicar que las nuevas generaciones están más familiarizadas, en general, con las tecnologías emergentes. En este sentido, se evidencia el cambio generacional. En consecuencia, se puede inferir que la educación a manos de las generaciones que actualmente se están formando podría incluir la IA y, en muchas ocasiones, este uso quizás implicaría menor formación que la que necesitan actualmente los docentes.

Con una perspectiva global de toda la prueba, la parte de «Producción oral» ha sido en la que ChatGPT ha podido «engañar» a más participantes, probablemente por la naturaleza de este ejercicio: más breve y variado en cuanto a formato. De hecho, la justificación que más se repetía era que las instrucciones no las solía usar la EOI. Por el contrario, es en la «Comprensión de textos escritos» donde ChatGPT se ha hecho más evidente entre los encuestados, sobre todo por la longitud de los textos, por la cual existía una mayor probabilidad de error y margen para encontrar diferencias entre el texto de la IA y el de la EOI.

Como se puede observar gráficamente en las figuras de cada sección, todas, a excepción de la 5 y de la 8 (correspondientes a la «Mediación escrita en español» y a la «Producción oral» respectivamente), tienen un porcentaje de respuestas correctas superior a la mitad, y el porcentaje de respuestas incorrectas en ningún caso es mayor, lo cual podría indicar que el rastro de ChatGPT no pasa desapercibido al ojo humano en la redacción de pruebas de evaluación. Sin embargo, se debe tener en cuenta que el número de «NS/NC» es relativamente alto, representando un tercio en las primeras figuras y una cuarta parte en las últimas.

Teniendo en cuenta las justificaciones expuestas anteriormente, se puede observar que ambos grupos de encuestados —los que respondieron correctamente y los que no— han detectado los mismos rasgos en muchas ocasiones (fallos, incoherencias, superficialidad, previsibilidad a causa de su lenguaje repetitivo y «tipo máquina», falta de actualidad y de acercamiento con el lector y, en un caso en concreto, desconocimiento del tipo de actividad que se le solicitaba). No obstante, la preferencia personal sobre la importancia de qué debe incluirse en una prueba de evaluación de alemán como lengua extranjera juega un papel importante en la decisión de los encuestados, ya que, mientras unos consideran ciertas características imprescindibles, otros las infravaloran, al mismo tiempo que sopesan otras como positivas y expresan que deberían incluirse en este tipo de exámenes.

Lo que está claro es que ChatGPT, por lo comentado constantemente por los participantes que sí lo han identificado, comete errores de vez en cuando, como el tuteo al examinando en varias ocasiones, de la misma forma que no siempre es coherente, incluso siendo este rasgo el fundamental para «engañar» y hacerse pasar por humano. No obstante, se ha observado que hay quienes asocian incoherencia con textos de la IA y, por ende, coherencia con textos de humanos; por lo tanto, la idea de que la coherencia podría hacer ver que se trata de un humano está meditada; sin embargo, en ocasiones, no es llevada a cabo con éxito.

La motivación (técnica) por la cual esta interfaz a veces peca de incoherente se desconoce en este caso, además de no ser el objetivo de la investigación. Sin embargo, no pasa desapercibida la evidente falta de conectores en las partes de «Producción escrita», ya que son precisamente estos los que sustentan un examen de esta índole y de este nivel. En el caso de

las respuestas erróneas que valoraban de manera positiva el uso de conectores por parte de la IA, era precisamente porque se trataba de conectores que se utilizan (casi) exclusivamente en el aula y no tanto en el habla de los nativos.

Otro inconveniente de ChatGPT es que, continuamente, le delata la formulación estereotipada y artificial en prácticamente todos sus textos, con una estructura muy similar en todos los casos. Esto lo ha detectado un gran número de participantes en varias ocasiones. La razón de ello, probablemente, sea que utiliza las mismas bases de datos para nutrirse y crear estos exámenes, aunque, independientemente de cuál sea el motivo exacto, la mayoría de los participantes han comprendido (y detectado), tras completar la encuesta, que la IA sigue una estrategia y que siempre es la misma o muy similar.

Adicionalmente, los temas de sus producciones también son de esa índole, sin referencias a la actualidad y sin capacidad de empatizar con el receptor, por lo que se convierte en un texto generalizado e impersonal. Como ya se comentaba en el primer apartado, ChatGPT no tiene esa empatía que un profesor sí puede mostrar, por lo que este aspecto es un punto en contra de la sustitución del profesorado por máquinas.

Asimismo, en una ocasión, algunos participantes, al dar una respuesta correcta, comentaban que dudaban de la posibilidad de cubrir la totalidad del enunciado de ChatGPT en las circunstancias de un examen y con el número de palabras que pide. Este hecho demuestra, de nuevo, la imposibilidad de la IA de ponerse en el lugar del examinando y meditar sus posibilidades, en este caso, más allá de su nivel.

Además, como se comentaba en una ocasión, ChatGPT parece no ser capaz de incluir elementos literarios y/o culturales en sus producciones. Por ello, se puede deducir que el factor humano es todavía una característica única y un elemento que lo diferencia de la máquina y, por tanto, lo hace insustituible.

Más aún, no siempre el texto de ChatGPT y las preguntas al respecto están relacionados, como es el caso de la ocasional diferencia de nivel del texto con el de las preguntas. Incluso, para completar los ejercicios, no se requiere una rigurosa comprensión del texto, es decir, que puede ser suficiente, en algunos casos, tan solo leerlo de manera superficial, sin requerir mucho esfuerzo por parte del examinando, elemento que, precisamente, se evalúa y, por ende, es de los más importantes en una prueba de lengua extranjera: la comprensión real del texto.

Por último, la IA no entiende el tipo de ejercicio del que se trata, lo cual se ha hecho evidente en ambas mediaciones, en las que ha solicitado al examinando otro tipo de ejercicio, además de no adecuarse a estándares que se dan por supuestos en estas circunstancias; por ejemplo, recordaba el tono en el que un estudiante debe escribir o el idioma en el que lo debe hacer, matices que también se examinan y no deberían incluirse en la prueba por regla general. En este sentido, ChatGPT hacía evidente la falta de experiencia en esta competencia.

El problema radica precisamente en lo que explicaba un encuestado que había respondido correctamente en la «Comprensión de textos escritos»: «el proceso de revisión, a veces, resulta más largo que el de búsqueda de textos ya terminados», por lo que cada usuario debe sopesar si prefiere beneficiarse de la IA y dedicar su tiempo a revisar lo que proporciona esta o si, por el contrario, prefiere invertir ese tiempo en buscar el texto por su cuenta (con lo que podrá controlar, por ejemplo, el vocabulario que incluye). Es una cuestión meramente personal, aunque el uso de ChatGPT parece no llegar a poder considerarse ventajoso para crear exámenes si el trabajo que implica es mayor al que se supone que ahorra. En cualquier caso, debe remarcarse que la producción de ChatGPT debe ser revisada por un humano, por todos los motivos que se exponen en el presente trabajo.

Con los resultados de este trabajo, también podríamos concluir que el uso de ChatGPT en el aula no es tan sencillo como los alumnos muchas veces suponen, sino que, más bien, se trata de una herramienta que requiere de tiempo (y no solo *a posteriori*), por lo que hay que sopesar si realmente vale la pena invertir en su uso —y si se trata realmente de una inversión—. Con todo esto, no se pretende afirmar que ChatGPT formula incorrectamente sus textos, ya que se puede observar en el propio trabajo que esto no es así: en la «Mediación escrita en español», varias respuestas correctas admiten que ChatGPT ha organizado mejor el texto que la EOI. De nuevo, se deben sopesar los rasgos que se consideran más oportunos para este tipo de pruebas. Un claro ejemplo es la mayor o menor formalidad que se le quiera dar al examen con tal de distanciarlo o acercarlo al examinando.

En consecuencia, se considera que actualmente el uso de ChatGPT para crear exámenes de lengua extranjera en un nivel B1 no es adecuada, ya que la

interfaz —todavía— comete errores que no pueden tolerarse por parte del autor de una prueba en la que se exige al examinando el mayor grado posible de precisión y conocimiento sobre la lengua. Tal vez sea la ignorancia acerca de las incorrecciones que comete ChatGPT, la que provoca un grado de confianza que se podría calificar como extremo por parte de muchos usuarios. Por lo que respecta a los participantes de este estudio, especialmente los profesores, se muestran cautelosos ante su suficiencia en el ámbito académico. De esta manera, se pone en duda que realmente estemos viviendo un cambio de paradigma crucial en el área de enseñanza de idiomas, puesto que el surgimiento de la IA implica un mayor esfuerzo humano (y, por ende, económico) que el que ya conocíamos antes de su existencia.

Así pues, en el marco de la hipótesis previa a la investigación, las TICS deben aún desarrollarse más con tal de poder confiar en ellas las pruebas de evaluación de alemán (nivel B1) y, así, reducir la labor docente de dicho idioma, debido a su incapacidad de crear exámenes sin —prácticamente— desaciertos y de manera rápida (se le tuvo que pedir mayor extensión y/o precisión en algunos ejercicios), especialmente en los apartados de examen que tan solo requieren participación pasiva por parte del examinando y, por tal razón, una mayor longitud y participación activa de ChatGPT. La refutación de la hipótesis inicial puede haberse dado a causa de la (excesiva) confianza que depositamos en las tecnologías en general y, en concreto, en ChatGPT, puesto a que se trata de una interfaz inteligente y que, en ocasiones, da la sensación de ser más coherente de lo que realmente es, si se analiza en profundidad.

Para concluir, siguiendo la reflexión de los dos participantes que, voluntariamente, comentaron que, en ocasiones, confiamos demasiado en la producción humana y demasiado poco en la capacidad que puede llegar a tener la IA de asemejarse a nosotros, no se pretende desprestigiar a ChatGPT ni a sus capacidades porque, además, es una tecnología emergente que, probablemente, mejore con el paso del tiempo. Sin embargo, con este trabajo se ha podido comprobar que, por ahora, esta interfaz debe realizar mayores avances para poder considerarse una alternativa seria a la labor que llevan a cabo los profesores de alemán como lengua extranjera en el ámbito universitario a la hora de realizar pruebas de evaluación.

Bibliografía

Alenizi, M. A. K., Mohamed, A. M., & Shaaban, T. S. (2023). Revolutionizing EFL Special Education: how ChatGPT is Transforming the Way Teachers Approach Language Learning. *Innoeduca*, *9*(2), 5–23. https://doi.org/10.24310/innoeduca.2023.v9i2.16774

Baker, P. (2023). *ChatGPT ™ for Dummies*. John Wiley & Sons, Inc.

Engelke, U., & Engelke, B. (2023). *ChatGPT - Mit KI in ein neues Zeitalter: Wie KI-Tools unser Leben und die Gesellschaft verändern* (1st ed.). mitp.

Fernández, Y. (2023, 1 junio). *ChatGPT Plus: qué es, diferencias con ChatGPT normal y cuánto cuesta esta inteligencia artificial*. Xataka. Recuperado el 25 de febrero de 2024, de https://www.xataka.com/basics/chatgpt-plus-que-que-caracteristicas-tiene-cuanto-cuesta-version-pago-esta-inteligencia-artificial

Fernández, Y. (2024, enero 10). *ChatGPT: qué es, cómo usarlo y qué puedes hacer con este chat de inteligencia artificial GPT*. Xataka. Recuperado el 25 de febrero de 2024, de https://www.xataka.com/basics/chatgpt-que-como-usarlo-que-puedes-hacer-este-chat-inteligencia-artificial

Hasanein, A. M., & Sobaih, A. E. E. (2023). Drivers and Consequences of ChatGPT Use in Higher Education: Key Stakeholder Perspectives. *European Journal of Investigation in Health, Psychology and Education*, *13*(11), 2599–2614. https://doi.org/10.3390/ejihpe13110181

Ichbiah, D. (2023). *ChatGPT : ¿quién eres?* Ediciones ENI. Recuperado el 27 de diciembre de 2023, de https://www.eni-training.com/portal/client/mediabook/home

***Información general de las enseñanzas de idiomas: Evaluación y Certificación* (s.f.).** EDUCAGOB. Portal del Sistema Educativo Español. Recuperado el 31 de diciembre de 2023, de https://educagob.educacionyfp.gob.es/ensenanzas/idiomas/informacion-general/evaluacion-certificacion.html

Kim, T. W. (2023). Application of artificial intelligence chatbots, including ChatGPT, in education, scholarly work, programming, and content generation and its prospects: a narrative review. *Journal of Educational Evaluation for Health Professions*, *20*, 38–38. https://doi.org/10.3352/jeehp.2023.20.38

Loos, E., Gröpler, J., & Goudeau, M.-L. S. (2023). Using ChatGPT in Education: Human Reflection on ChatGPT's Self-Reflection. *Societies (Basel, Switzerland)*, *13*(8), 196. https://doi.org/10.3390/soc13080196

Marco Común Europeo de Referencia para las Lenguas: Aprendizaje, Enseñanza, Evaluación (2000). En *Centro Virtual Cervantes* (NIPO: 176-02-187-X). ANAYA.

Mondal, H., Marndi, G., Behera, J., & Mondal, S. (2023). ChatGPT for Teachers: Practical Examples for Utilizing Artificial Intelligence for Educational Purposes. *Indian Journal of Vascular and Endovascular Surgery*, *10*(3), 200–205. https://doi.org/10.4103/ijves.ijves_37_23

OpenAI (2023). ChatGPT (versión del 31 de diciembre). https://chat.openai.com/chat

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Sabzalieva, E., & Valentini, A. (2023). *ChatGPT e inteligencia artificial en la educación superior: guía de inicio rápido* (ED/HE/IESALC/IP/2023/12). https://unesdoc.unesco.org/ark:/48223/pf0000385146_spa

We asked ChatGPT: How does artificial intelligence change teaching? (2023). En *CE Noticias Financieras* (English ed.). ContentEngine LLC, a Florida limited liability company.

Your German SAT questions were written by ChatGPT (2023). En *CE Noticias Financieras* (English ed.). ContentEngine LLC, a Florida limited liability company.

Anexo I: Examen de la EOI de La Rioja (junio, 2022) (B1)
https://drive.google.com/drive/folders/10Gk4ntYK3hAAIXZ8Vt8JUAMH-johxwLJw

Anexo II: Cuestionario al profesorado de alemán
https://docs.google.com/forms/d/e/1FAIpQLScxN0_78y6fcFLSD3W3etruLSM_TysUXWd8Cd2t49zF3FELAg/viewform?usp=pp_url

Anexo III: Cuestionario al alumnado de alemán
https://docs.google.com/forms/d/e/1FAIpQLScYmTNTSCdKRwUr6j4qlC2aL6cMunTDJJBlVAuz-PYW3Ck4Wg/viewform?usp=pp_url