EVOLUCIÓN DE LA EVALUACIÓN DE L2: ESCENARIOS ELECTRÓNICOS E IA: UN ESTUDIO EXPLORATORIO

Adriana Merino
Princeton University (Estados Unidos)
amerino@princeton.edu

Resumen

Este capítulo explora el potencial de la Inteligencia Artificial (IA) para mejorar la evaluación de idiomas a través de escenarios simulados que abordan de modo integral los modos de comunicación interpretativo, interpersonal y presentacional. Se centra en la evaluación formativa, donde los estudiantes demuestran habilidades en contextos prácticos, apoyados por chatbots, avatares y tecnología adaptativa para evaluar competencias lingüístico-comunicativas en un entorno seguro.

El estudio[1], aún en proceso de desarrollo, tiene como objetivo evaluar la eficacia de la IA en la evaluación formativa de estudiantes de segundas lenguas (L2), su adaptabilidad en diversos escenarios, comparar su rendimiento con evaluadores humanos, así como entender cómo los estudiantes perciben sus logros y objetivos en relación con este método de evaluación.

1. INTRODUCCIÓN

La evaluación educativa ha sido tradicionalmente considerada como una herramienta de medición para determinar la eficacia y la competencia del proceso de enseñanza-aprendizaje. Sin embargo, en el contexto de la Educación Superior, y particularmente en la adquisición de un segundo idioma, la evaluación se transforma en un instrumento de valoración que no solo mide, sino que también influye sobre las capacidades, percepciones y roles de los estudiantes como aprendices del idioma (Black & William 1998).

En las últimas décadas, el campo de la evaluación para el aprendizaje ha experimentado una transformación significativa debido a la incursión de las tecnologías de la información y comunicación. Las prácticas tradicionales están cediendo espacio a herramientas y enfoques innovadores que buscan responder eficazmente a las necesidades cambiantes del siglo XXI.

En este marco de cambio constante, las plataformas digitales y, de manera más reciente, la Inteligencia Artificial (IA), han emergido como protagonistas en el rediseño de la evaluación de la competencia lingüístico-comunicativa y pragmática en la enseñanza de lenguas extranjeras (L2). Este capítulo ofrece una visión hacia el futuro, revisando las tendencias de la evaluación de los últimos años y buscando establecer un panorama actualizado para orientar futuras investigaciones y prácticas en el ámbito de la evaluación de lenguas extranjeras en el marco digital y de la IA.

2. EVALUACIÓN FORMATIVA Y LA EVALUACIÓN ORIENTADA AL APRENDIZAJE

2.1. Definiciones

Dada la utilización de diversas nomenclaturas a los diseños de evaluación emergentes, es esencial una clarificación terminológica. En primer lugar, defino lo que se entiende por evaluación formativa y la evaluación orientada al aprendizaje (EOA).

La evaluación formativa es un proceso continuo que recopila información sobre el aprendizaje de los estudiantes para mejorarlo. A diferencia de la evaluación sumativa, que concluye un período de aprendizaje, la formativa es progresiva y se enfoca en el proceso educativo, evaluando el rendimiento del estudiante durante todo el periodo. Teóricos como Scriven (1967) y Bloom (1969), ya en la década de los 60, abogaban por un método que comprobara la eficacia de un plan de estudios y que pudiera modificarse en caso de necesidad de mejoras. Esta perspectiva comenzaba a subrayar la importancia de una evaluación integral, continua y centrada en el desempeño real a fin de potenciar el desarrollo integral del estudiante.

En la enseñanza de lenguas extranjeras, la evaluación formativa detecta los aspectos lingüísticos dominados por el alumno y aquellos que requieren mejora. Esto posibilita que el profesor personalice su enseñanza según las necesidades de cada estudiante. La retroalimentación específica, constructiva y oportuna

[1] Esta investigación se enmarca en el proyecto de Pedagogía Digital e Inteligencia Artificial en L2, patrocinado por *McGraw Center for Teaching and Learning, Princeton University* (USA), cuyos investigadores principales son: Adriana Merino, PhD. *Senior Lecturer in Spanish* y Ben Johnston, *Senior Educational Technologist, Digital Learning and Design*.

es esencial en este enfoque, facilitando a los alumnos la superación constante. Además, al fortalecer la motivación y la autoestima, la evaluación formativa puede incrementar la confianza y el compromiso de los estudiantes con su aprendizaje. Técnicas como observaciones en clase, autoevaluaciones y coevaluaciones, que incluyen retroalimentación mutua entre compañeros, son cruciales para identificar fortalezas y debilidades.

Por otra parte, a principios del siglo XXI, surgió la Evaluación Orientada al Aprendizaje (EOA), término acuñado por Carless (2003), como un modelo de evaluación centrado en el estudiante. Esta metodología prioriza el desarrollo de habilidades y competencias sobre la mera medición de conocimientos. Originada en Europa, Canadá, EE.UU. y la región Asia-Pacífico, la EOA se presenta como un método de evaluación alternativo que considera el aprendizaje como un proceso continuo y fundamental en un modelo de enseñanza centrado en el alumno. Autores como Dochy, Segers y Dierick (2002) sugirieron que las universidades deberían transitar de culturas evaluativas basadas en exámenes, por lo general sumativos, a culturas que fomenten la evaluación *para* el aprendizaje. La EOA evolucionó tanto de la evaluación sumativa como formativa, integrando tres componentes: evaluación *para* el aprendizaje, evaluación *como* aprendizaje y evaluación *del* aprendizaje.

En la EOA, la evaluación cumple dos funciones principales. Primero, actúa como una herramienta *de* aprendizaje donde la metacognición del estudiante es clave, ayudando en la reflexión y el seguimiento del progreso para orientar objetivos futuros. Segundo, las evaluaciones están diseñadas *para* el aprendizaje, proporcionando orientación tanto a estudiantes como a instructores sobre los niveles actuales de comprensión. Esta forma de evaluación es significativa dado que se alinea estrechamente con el currículo y promueve la reflexión, impulsando habilidades de pensamiento de orden superior. Asimismo, fomenta un conocimiento profundo y duradero más allá de la memorización superficial. Para su efectividad, es vital que los resultados de la evaluación se presenten de manera clara, permitiendo que tanto los docentes como los discentes comprendan los pasos a seguir. Sin un contexto claro, las calificaciones genéricas no proporcionan la guía necesaria para decisiones pedagógicas efectivas. Por lo tanto, es imprescindible poder proyectar la aplicabilidad futura de esta información.

En suma, aunque tanto la Evaluación Formativa como la Evaluación de Orientación al Aprendizaje (EOA) están centradas en el aprendizaje del estudiante, difieren en su enfoque: la primera se centra en el proceso de aprendizaje, mientras que la segunda prioriza el desarrollo de habilidades y competencias. En ambas modalidades, se hace énfasis en la "alineación constructiva" que sostiene que la evaluación, el currículo, los materiales didácticos, los resultados esperados y las metodologías de enseñanza deben estar alineados y estrechamente interconectados.

2.2. La Evaluación de Desempeño Integrada (EDI)

La Evaluación de Desempeño Integrada (EDI), desarrollada por la American Council on the Teaching of Foreign Languages (ACTFL) en la década de 1990, representa una aplicación práctica de la Evaluación Orientada al Aprendizaje (EOA). Esta metodología propone una evaluación frecuente, holística y dinámica de las competencias del aprendiz en diversos contextos y modalidades comunicativas, con el propósito de potenciar el aprendizaje mediante retroalimentación constructiva. Abarca tres modos comunicativos esenciales que se interrelacionan de manera coherente en un enfoque de multiliteracidades: el modo interpretativo, orientado hacia la comprensión profunda del contenido; el modo interpersonal, que resalta la interacción y el flujo de información; y el modo presentacional, que se dirige hacia la elaboración y exposición de contenidos frente a un público. Se propone en este trabajo una cuarta dimensión que se integra a los otros modos: la reflexiva. Esta última incentiva la autoevaluación y reflexión estudiantil sobre su aprendizaje y desempeño, promoviendo una autorregulación y metacognición más robusta en el ámbito educativo. Juntas, estas dimensiones conforman una evaluación comprensiva que trasciende la simple cuantificación y se centra en el crecimiento sostenido del estudiante.

2.3. Evaluaciones basadas en escenarios (EBE)

La Evaluación Basada en Escenarios (EBE), como una parte integral e integradora del proceso de enseñanza-aprendizaje de L2, y alineada con la EOA y EDI, representa una evolución significativa, enfocándose en un aprendizaje que es tanto auténtico como aplicable. Se inspira en la experiencia orientada a la acción, proveyendo a los estudiantes de situaciones que requieren la utilización conjunta de conocimientos y competencias multimodales para abordar diversos desafíos comunicativos. Este enfoque innovador[2] se

[2] La metodología de evaluación mediante escenarios, aunque no es nueva en la evaluación del aprendizaje, ha cobrado relevancia en años recientes, especialmente en la evaluación de segundas lenguas (L2). Richards y Lockhart (1994), fueron pioneros en proponer este enfoque, centrado en la creación de situaciones de comunicación auténticas para que los estudiantes demuestren sus habilidades en L2. Posteriormente, Bardovi-Harlig y Dörnyei (1998), y Bygate (2009), expandieron este concepto, enfatizando la importancia de escenarios y actividades de aprendizaje que permitan a los estudiantes aplicar sus habilidades en L2 en diversos contextos. Este enfoque holístico y comunicativo ha ganado popularidad por su efectividad en reflejar situaciones reales de uso del lenguaje.

distingue por crear entornos de aprendizaje que reflejan situaciones de la vida real, permitiendo a los estudiantes aplicar sus habilidades lingüístico-comunicativas en contextos prácticos y significativos.

La Evaluación Basada en Escenarios (EBE) se fundamenta en tres pilares: la interacción auténtica, el uso holístico de habilidades lingüísticas y la reflexión metacognitiva. La interacción auténtica se centra en capacitar a los estudiantes para comunicarse en contextos realistas, superando las limitaciones del aula tradicional. Por su parte, el uso holístico de habilidades lingüísticas une lectura, escritura, expresión oral, y comprensión auditiva y visual en un enfoque integrado, donde el estudiante es un protagonista activo y usuario esencial del idioma. La reflexión metacognitiva, inspirada en teorías como la de Nelson y Narens (1990), sobre evaluación en contextos auténticos, insta a los estudiantes a analizar su propio aprendizaje, identificando fortalezas y áreas de mejora aplicables en situaciones reales de comunicación.

En la EBE, la tecnología es vital no solo para simular entornos de aprendizaje realistas, sino también como herramienta pedagógica que enriquece la experiencia educativa. Facilita la resolución de problemas y promueve el pensamiento crítico, la toma de decisiones y el uso efectivo de habilidades comunicativas. La integración tecnológica en la EBE aporta flexibilidad, creatividad y relevancia a los escenarios de aprendizaje, incrementando el impacto educativo y fomentando la autonomía del estudiante. Asimismo, la resolución de problemas representa una estrategia constructivista clave que motiva a los estudiantes a adquirir conocimientos y habilidades pertinentes de modo activo, basado en la práctica.

3. MIRANDO AL FUTURO: ¿HACIA DÓNDE VA LA EVALUACIÓN DE L2?

Como hemos apuntado en los apartados anteriores, las últimas tendencias en la evaluación de idiomas extranjeros reflejan un cambio significativo hacia enfoques más constructivistas, centrados en la autenticidad, la practicidad y las habilidades performativas. Se evidencia una transición hacia una evaluación más colaborativa y ajustable en plataformas digitales que potencian la implicación de los estudiantes, alineándose con sus necesidades y preferencias. Se enfatiza, asimismo, la multiliteracidad, la resolución de problemas y la capacidad de los estudiantes para aplicar lo aprendido en situaciones reales, donde el docente se transforma en mediador, y facilita la autonomía del discente. En este nuevo paradigma, la inteligencia artificial (IA) emerge como una herramienta vital, no solo para proporcionar retroalimentación inicial a los estudiantes, sino también para asistir a los docentes en la evaluación inicial de las tareas.

3.1. La inteligencia artificial en la evaluación de idiomas extranjeros

En la confluencia de la era digital y la Inteligencia Artificial (IA), este estudio se ha centrado en explorar el potencial de las aplicaciones basadas en IA, como chatbots, generadores de voz, texto y avatares, en la evaluación de la competencia comunicativa y sociocultural de los estudiantes de segundo idioma. Partimos de la premisa de que estas herramientas no solo proporcionan un método de evaluación más dinámico y atractivo, sino que también facilitan un camino hacia una enseñanza más personalizada y efectiva. Un objetivo clave de esta investigación ha sido determinar si, en evaluaciones inmersivas basadas en escenarios, la IA puede asistir en la evaluación del desempeño multimodal del aprendiz en tiempo real. La IA podría enriquecer las prácticas evaluativas, promoviendo un enfoque más holístico y adaptativo para medir y mejorar las habilidades lingüístico-comunicativas. Esta investigación representa un paso adelante en la comprensión y aplicación de tecnologías avanzadas en el ámbito de la enseñanza de lenguas extranjeras.

3.2. La plataforma digital para la evaluación

Para investigar el potencial de la Inteligencia Artificial (IA) en la implementación de evaluaciones electrónicas basadas en escenarios, hemos desarrollado una plataforma digital interactiva y multimodal, destinada tanto a la evaluación como al aprendizaje. Esta plataforma presenta una variedad de escenarios (ver Figura 1) que asignamos a estudiantes de español de niveles principiante e intermedio, siguiendo un calendario predefinido, aproximadamente cada tres o cuatro semanas de instrucción. Comunicamos a los estudiantes que el propósito de estas evaluaciones basadas en escenarios auténticos es medir su capacidad para aplicar conocimientos y habilidades en situaciones reales. Se enfrentarán a un escenario que simula un desafío o situación auténtica, siendo su rendimiento en esta simulación el principal criterio de evaluación.

Para asegurar la transparencia del proceso, informamos a los estudiantes que recibirán tanto retroalimentación (feedback) como orientación para su desarrollo futuro (feedforward) a través de una combinación de evaluación humana (por parte del instructor) y evaluación asistida por IA, como Chat GPT. Se espera que el estudiante comprenda la situación presentada, identifique problemas pragmático-comunicativos y reflexione sobre ideas o situaciones realistas, lo que implica el uso de habilidades cognitivas de orden superior.

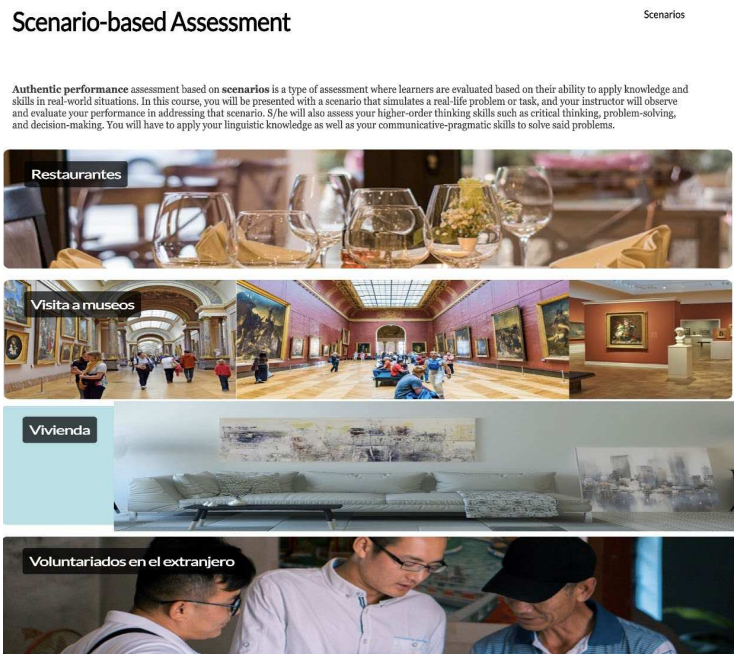

Figura 1. Una captura de pantalla de la plataforma digital para evaluaciones. Princeton University. USA

En nuestra plataforma, recreamos situaciones cotidianas como una cena en un restaurante con amigos que tienen restricciones dietéticas, la elección de una vivienda adecuada para un grupo con distintas necesidades, o una visita guiada a un museo, subrayando así la autenticidad y relevancia de estos escenarios. Los estudiantes tienen la oportunidad de aplicar sus conocimientos en un formato accesible y controlado. Además, la integración de tecnología adaptativa e Inteligencia Artificial personaliza las respuestas, y la interacción con avatares y la conversión de voz a texto hacen la experiencia de aprendizaje más significativa y adaptada a cada individuo.

3.2.1. Muestra: Voluntariado en el extranjero

A modo de ejemplo, consideremos el siguiente escenario de nuestro estudio exploratorio llevado a cabo en la Universidad de Princeton durante el otoño de 2023: "Voluntariado en el Extranjero", donde los estudiantes atraviesan una secuencia de 4 o 5 pasos (ver Figura 2) que replica los procesos propios de los programas de voluntariado. En primer lugar, los estudiantes deben decidir su destino y el tipo de voluntariado, según sus intereses académicos y personales. Esto incluye la selección de un país y el alojamiento preferido. Tras esto, el/la estudiante debe solicitar un puesto como voluntario(a) mediante la presentación de una solicitud formal que incluye una declaración personal. Una vez aceptado(a), el siguiente paso implica asistir a una entrevista, donde el/la estudiante intenta convencer al entrevistador de su idoneidad para el rol. La fase final conlleva promocionar su experiencia a través de una presentación ante sus compañeros universitarios, detallando el proceso y los beneficios de tal experiencia.

3.2.1.1. Explicación

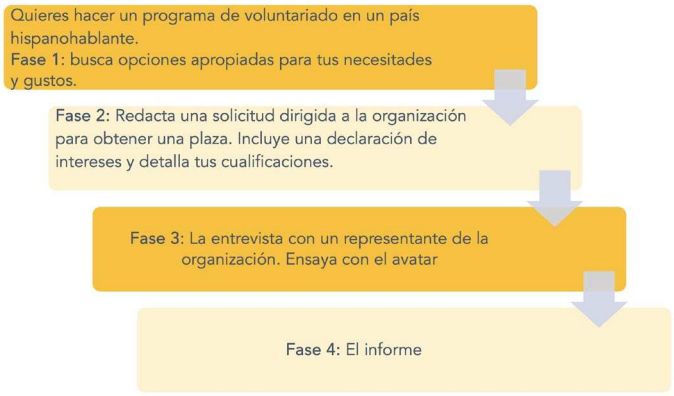

Figura 2. Fases del escenario. Voluntariado en el extranjero

Fase 1: En esta etapa, los estudiantes emplean el Modo Interpretativo de comunicación, que implica entender significados en distintos formatos como programas de radio o TV, textos escritos, fragmentos de películas, etc. En el caso de este escenario (ver Figura 3), el enfoque radica en la interpretación de información

proporcionada por tres buscadores impulsados por inteligencia artificial: Bard, Bing y Phind[3]. Los alumnos utilizan habilidades de lectura y comprensión visual, y en el caso de Bard, también auditivas. Su tarea es investigar sobre oportunidades de voluntariado en países hispanohablantes, filtrando programas según su carrera, intereses académicos y personales. Este proceso requiere leer, interpretar y clasificar información, y luego tomar decisiones basadas en sus hallazgos e intereses, justificando sus elecciones. Los procesos cognitivos involucrados en la comprensión lectora en un idioma extranjero son análogos a los de la habilidad auditiva, involucrando competencias gramaticales, discursivas, sociolingüísticas y estratégicas.

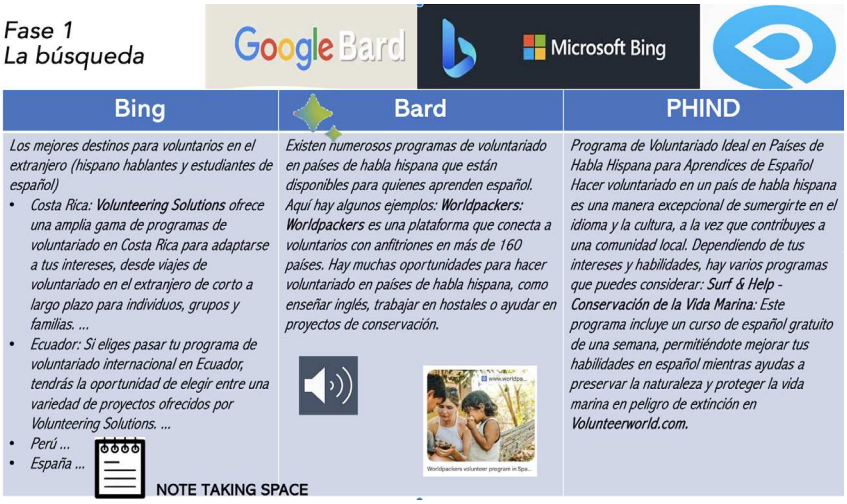

Figura 3. Fase 1: La búsqueda

Dado que se trata de una evaluación de 90 minutos, les proporcionamos la búsqueda pre-digerida. Sin embargo, si se tratara de una evaluación de mayor duración, el docente puede requerir que los alumnos realicen la búsqueda por sí mismos. Luego, deben hacer un cuadro comparativo con los *pros y contras* de cada opción, elegir y justificar su elección detalladamente.

Fase 2: Esta fase requiere el uso del Modo Interpersonal que implica comunicación escrita y oral bidireccional con negociación activa de significados. En este caso, el alumno escribe una carta de presentación y una declaración de interés a la organización de su elección. En razón de esta característica, está involucrada la habilidad de escribir. En este caso, la producción de escritura no requiere utilización de IA, aunque el docente sí la puede utilizar para la evaluación (ver Apartado 3.1.)

Fase 3: La entrevista. Esta fase aplica una variante del Modo Interpersonal de comunicación, centrada en la interacción oral bidireccional que facilita una negociación activa de significados. Aquí, el estudiante se involucra en una práctica de entrevista con un avatar de IA, como Revoicer o Hey Gen. Este avatar (ver Figura 4) formula preguntas habituales en entrevistas de voluntariado, tales como "¿Por qué quieres hacer este voluntariado?", "¿Por qué elegiste este país?", "¿Has participado previamente en un voluntariado similar?" y "¿Prefieres hospedarte en una casa de familia o en un hostal?". El alumno debe responder a cada pregunta en un tiempo determinado, grabando sus respuestas. Gracias a la tecnología adaptativa, el avatar ajusta sus preguntas basándose en las respuestas previas del alumno. En esta actividad, el estudiante pone a prueba y mejora sus habilidades de comprensión auditiva, así como sus competencias en expresión oral y fluidez.



Figura 4. Fase 3: La entrevista

[3] Bing, Bard (recientemente renombrada Gemini) y Phind son motores de búsqueda que integran inteligencia artificial. Gemini (ex Bard), de Google, es un modelo de lenguaje conversacional para búsqueda en Internet y generación de texto. Bing, de Microsoft, evolucionó desde un motor tradicional a uno con IA, ofreciendo búsquedas avanzadas. Phind se centra en consultas técnicas y de programación, proveyendo respuestas detalladas para desarrolladores. Estos buscadores ejemplifican la transformación de la búsqueda de información en la era de la IA.

Fase 4: Regreso a Casa. En esta etapa, se solicita al estudiante realizar una presentación ante sus compañeros universitarios en la oficina de estudios y viajes internacionales, con el objetivo de promover experiencias de voluntariado. Esta fase emplea el Modo Presentacional de comunicación, caracterizado por ser unidireccional y sin interacción directa para la negociación de significados con la audiencia. Este modo exige habilidades específicas tanto en expresión oral como escrita. El estudiante debe demostrar competencias en la organización clara de ideas, la capacidad de sintetizar experiencias y transmitirlas de manera efectiva, así como el uso adecuado del lenguaje para captar y mantener el interés de la audiencia. Además, se requiere habilidad para adaptar el contenido a un contexto formal y académico, asegurando que la presentación sea informativa, atractiva y culturalmente relevante.

Como se ve en esta muestra de Evaluación Electrónica Basada en Escenarios (EeBE) que proponemos, compuesta por cuatro fases distintas, representa una estrategia educativa que destaca la evaluación auténtica, simulando situaciones reales y desafíos cotidianos que el estudiante podría enfrentar. Este enfoque no solo valida la relevancia del aprendizaje sino que también lo contextualiza, haciéndolo más significativo y aplicable.

Un aspecto clave de esta metodología es la integración de las multiliteracidades en el proceso de evaluación de una segunda lengua (L2). Esta integración es fundamental, ya que reconoce que las competencias lingüísticas no funcionan de manera aislada sino que están interconectadas y se enriquecen mutuamente. Al abordar la evaluación (así como el aprendizaje) de un idioma desde múltiples perspectivas y habilidades, se fomenta una comprensión más profunda y versátil del lenguaje.

Además, cada fase de esta evaluación incorpora el Modo Reflexivo, impulsando al alumno a ejercitar habilidades críticas de pensamiento y resolución de problemas. Este enfoque reflexivo no solo mejora la comprensión y el aprendizaje del contenido, sino que también promueve el desarrollo de una mayor conciencia de sí mismo y del entorno. Estas habilidades son esenciales no solo para el aprendizaje de idiomas, sino también para la formación integral del individuo, preparándolo para navegar y contribuir efectivamente en un mundo cada vez más globalizado y complejo.

4. EVALUACIÓN ASISTIDA POR IA Y RETROALIMENTACIÓN

Los sistemas de IA también pueden proporcionar retroalimentación en tiempo real a los estudiantes, lo que permite una evaluación y un aprendizaje más rápidos. A continuación, se muestra la capacidad de IA para corregir y dar feedback de la producción escrita. Creemos que la IA puede ser especialmente valiosa en la evaluación de habilidades orales como la pronunciación, donde la retroalimentación inmediata puede ayudar a corregir errores a medida que ocurren. Este será el siguiente paso de este estudio.

4.1. La IA como herramienta de evaluación

A fin de comparar el rendimiento de Chat GPT con los evaluadores humanos en la corrección de errores en distintas categorías, así como las capacidades y potencial de Chat GPT como herramienta evaluadora, se proporcionaron a GPT cinco composiciones de distinta longitud. Se le solicitó a GPT identificar errores en distintas categorías, como vocabulario, sintaxis, colocaciones, tiempos y formas verbales, orden de palabras y concordancia. Además, se le pidió que evaluara el contenido según se ajustaba a las instrucciones o si se desviaba del tema propuesto. En cuanto al feedback, después de detectar y resaltar los errores en negrita, se instruyó a GPT para que proporcionara comentarios explicativos sobre cada error y, posteriormente, ofreciera una versión corregida y mejorada de la frase o enunciado incorrecto.

Al comparar las correcciones realizadas por GPT con las evaluaciones de revisores humanos (ver Figura 5), se observó una alta coincidencia en los puntajes, según los cálculos de la prueba T que llevamos a cabo. GPT identificó los siguientes errores: 8 en vocabulario, 15 en concordancia, 10 en tiempos verbales, 5 en formas verbales y 4 en colocaciones. Por su parte, cinco profesores humanos detectaron 9 errores en vocabulario, 14 en concordancia, 9 en tiempos verbales, 6 en formas verbales y 3 en colocaciones.

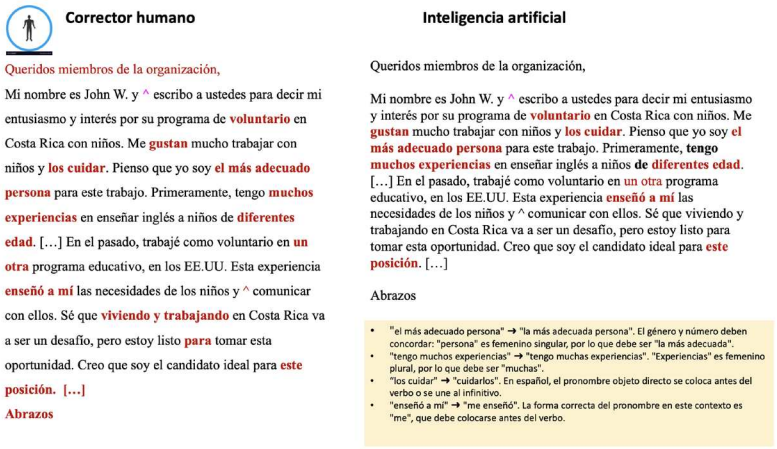

Figura 5. Muestra de composiciones analizadas y corregidas por Chat GPT y profesores.

4.2. Evaluador humano *vs*. evaluador de IA

Como se dijo arriba, la revisión de los textos en español realizada tanto por la inteligencia artificial (ChatGPT) como por los revisores humanos mostró una alta coincidencia en la detección de varios tipos de errores, pero también reveló diferencias significativas en ciertas áreas, especialmente en errores pragmáticos y de estilo. A continuación, se detallan los tipos de errores identificados y las diferencias en las revisiones.

Los resultados revelan una notable congruencia entre las correcciones realizadas por Chat GPT y las evaluaciones de los correctores humanos, destacando la eficacia de la IA en la identificación y corrección de aspectos clave como vocabulario, sintaxis, tiempos verbales, colocaciones y concordancia. La similitud en las correcciones sugiere que Chat GPT posee una capacidad avanzada para detectar y rectificar errores comunes en textos escritos, alineándose efectivamente con los estándares humanos de corrección lingüística. Esta eficiencia notable en detectar y corregir errores técnicos o estructurales se debe a su entrenamiento en extensos volúmenes de datos textuales que le permite aprender patrones lingüísticos y reglas gramaticales

Errores comunes identificados:

- Concordancia: ambos, la IA y el evaluador humano, identificaron errores de concordancia, como la falta de coincidencia en número o género entre sustantivos y adjetivos.
- Vocabulario: errores en la elección de palabras adecuadas o expresiones idiomáticas se señalaron correctamente.
- Uso de pronombres: errores en la colocación y uso de pronombres directos e indirectos fueron notados por ambos revisores.
- Forma y Tiempo Verbal: inconsistencias en la conjugación de verbos y el uso de tiempos verbales fueron detectadas.

Por otro lado, el estudio revela que los evaluadores humanos superan a las evaluaciones por IA en aspectos más sutiles y complejos, como el estilo o el tono, donde la sensibilidad y la comprensión humana juegan un papel importante en sus apreciaciones. La investigación, por lo tanto, revela un panorama mixto en cuanto a la eficacia de Chat GPT y los evaluadores humanos, resaltando tanto las potencialidades como las limitaciones de la inteligencia artificial en el campo de la corrección lingüística.

Errores detectados principalmente por el revisor humano:

- Errores pragmáticos contextuales: aspectos como la formalidad en la salutación (“Estimados” en lugar de “Queridos”) y el cierre (“Un saludo cordial” en lugar de “Abrazos”) son ejemplos de errores pragmáticos relacionados con el contexto y el tono de la comunicación, más fácilmente identificados por humanos.
- Errores de estilo: Detalles como el uso de dos puntos en lugar de comas después de la salutación en español son aspectos de estilo de escritura que a menudo requieren un conocimiento cultural y lingüístico más profundo.
- Posibles diferencias en la evaluación del contenido. Aunque la IA es eficaz en la corrección técnica, puede haber diferencias en la evaluación del contenido, especialmente en términos de relevancia y ajuste a las instrucciones. Los humanos podrían tener una mejor comprensión del contexto y los matices que requiere evaluar si un contenido se ajusta a las instrucciones dadas.

En los hallazgos también se destaca la consistencia de Chat GPT en las correcciones y la variabilidad en las correcciones humanas, reflejando un enfoque más personalizado y contextual en la evaluación. Chat GPT

no solo identifica errores, sino que también proporciona feedback, esto es, explicaciones detalladas y versiones mejoradas, facilitando el aprendizaje y la mejora del usuario. Además, se observa una complementariedad entre la IA y los evaluadores humanos: mientras la IA ofrece una revisión técnica y rápida, los humanos aportan una comprensión más profunda del contexto y los matices del lenguaje. Estas diferencias sugieren que la combinación de ambos enfoques puede ser especialmente eficaz para una evaluación exhaustiva y matizada.

4.3. Calificaciones numéricas y rúbricas

En cuanto a la calificación numérica, el estudio exploró también las diferencias entre la evaluación de composiciones escritas hecha por Chat GPT y por los evaluadores humanos. Se usó una rúbrica estandarizada para calificar cinco composiciones tanto por Chat GPT como por evaluadores humanos. Los resultados mostraron que Chat GPT tendía a dar calificaciones más bajas que los humanos, lo que indica que, aunque la IA es eficaz en un análisis basado en criterios técnicos, los humanos aportan una comprensión más contextual y matizada (ver Figura 6)

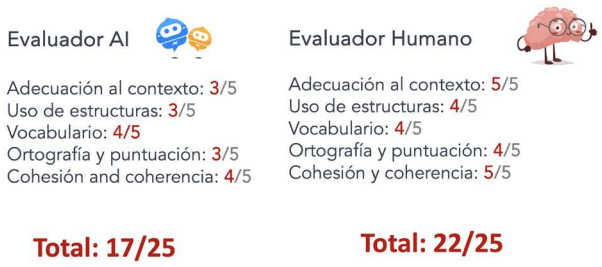

Figura 6. AI vs Evaluador humano en la calificación numérica

A modo de explicación podemos decir que Chat GPT, al ser una herramienta basada en inteligencia artificial, se enfoca en aspectos técnicos y objetivos del texto, como la gramática, ortografía y estructura. Este enfoque más analítico de la IA puede llevar a una evaluación más rigurosa y, posiblemente, más crítica, reflejada en calificaciones más bajas. Por su parte, los evaluadores humanos pueden considerar aspectos más subjetivos y contextuales, como la creatividad, el estilo, la intención del autor, la relevancia del contenido y hasta la resonancia emocional del texto. Esta capacidad para "contextualizar" puede resultar en una apreciación más amplia de la calidad del texto, llevando a calificaciones más altas. Otro aspecto a destacar también en la evaluación numérica tiene que ver con la consistencia *vs*. la subjetividad, mientras que Chat GPT ofrece una consistencia en la evaluación, los humanos presentan una variedad de perspectivas. Esta subjetividad humana puede ser tanto una fortaleza como una limitación, dependiendo del objetivo de la evaluación. Es posible que Chat GPT y los evaluadores humanos estén utilizando diferentes criterios para asignar calificaciones, incluso si se les proporciona la misma rúbrica. La interpretación de los criterios puede variar significativamente, especialmente en aspectos cualitativos.

En resumen, la investigación sugiere que mientras Chat GPT es eficiente en analizar aspectos técnicos y objetivos de las composiciones, los evaluadores humanos aportan una comprensión más profunda y contextual, lo cual es crucial en la evaluación de textos más subjetivos y creativos. Esto destaca la importancia de una evaluación equilibrada que incorpore tanto perspectivas técnicas como contextuales. La IA puede proporcionar una primera capa de análisis basado en criterios objetivos, mientras que los docentes pueden aplicar su comprensión contextual y subjetiva para ajustar y finalizar las calificaciones. De esta manera, la IA puede aligerar la carga de trabajo de los docentes y permitirles concentrarse en aspectos más profundos y matizados de la evaluación de textos.

5. PERCEPCIÓN DE LA EVALUACIÓN POR PARTE DE LOS ALUMNOS

Este estudio exploratorio también abarca la perspectiva del alumno, enfatizando la evaluación como un proceso colaborativo y no adversarial. Para comprender cómo los alumnos perciben la evaluación performativa y formativa impulsada por la IA, se realizó una encuesta piloto con 15 estudiantes. Esta encuesta se centró en aspectos como los beneficios percibidos de la evaluación, la superación de barreras comunicativas, mejoras en la expresión escrita, revisión de vocabulario y estructuras gramaticales, y la autenticidad de la tarea. La inclusión de las opiniones de los alumnos es crucial para comprender el impacto completo de la evaluación asistida por IA en el proceso educativo.

5.1. Resultados

Los resultados arrojaron conclusiones valiosas sobre la eficacia de estos métodos en el ámbito lingüístico y comunicativo. El 73% de los alumnos encuestados consideró que la simulación de escenarios realistas, donde se les sitúa en un contexto auténtico, no solo es beneficioso para su aprendizaje general, sino que poten-

cia su capacidad de comprensión, aplicación del conocimiento y de consolidación de aspectos gramaticales y léxicos, dándoles la oportunidad de aplicar de manera práctica el lenguaje en contextos reales.

En cuanto a los beneficios específicos que obtuvieron en su aprendizaje (Ver Figura 7), un 60% destacó la mejora en habilidades prácticas y de resolución de problemas, mientras que el 40% restante subrayó la adquisición de habilidades interpersonales y comunicativas. Estos datos nos indican que, más allá de los conocimientos teóricos, los escenarios auténticos permiten a los estudiantes desarrollar habilidades vitales para la vida real y profesional.

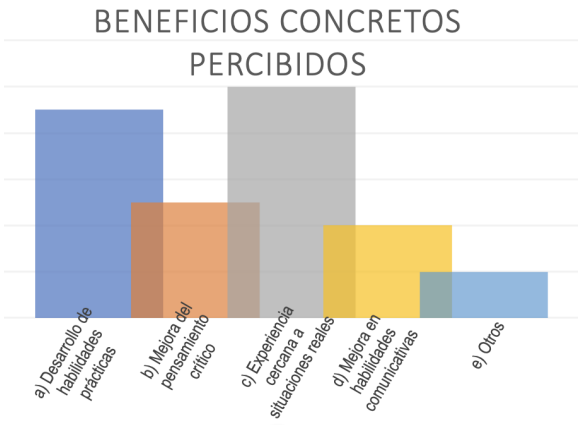

Figura 7. Beneficios concretos percibidos por los estudiantes de las pruebas por escenarios.

Al preguntarles acerca de las interacciones con avatares en las evaluaciones (ver Figura 8) el 53% caracterizó estas interacciones como "enriquecedoras" y "motivadoras", mientras que el 47% las describió como "neutras". Una de las ventajas más notables de las evaluaciones basadas en escenarios auténticos con avatares es la reducción de la ansiedad. El 53% de los estudiantes mencionó que hablar con avatares generó menos ansiedad que interactuar con humanos. Esto les permitió centrarse más en el contenido y la estructura de su comunicación, en lugar de preocuparse por cometer errores o ser juzgados. Es importante mencionar que ningún estudiante las calificó como negativas. Esto podría señalar que, si bien la mayoría encontró valor en la interacción con avatares, aún hay espacio para mejorar la experiencia y hacerla más significativa para todos.



Figura 8. Interacción con avatares

Además, un aspecto fundamental que resaltaron los estudiantes fue el feedback recibido acerca de su mejora en la performance oral y escrita. Un 60% señaló que las correcciones y comentarios proporcionados por el sistema les ofreció una visión clara sobre sus áreas de fortaleza y aquellas que necesitan mejorar. Este tipo de retroalimentación inmediata y personalizada es esencial para el proceso de aprendizaje de un idioma.

6. CONCLUSIONES

La incorporación de la Inteligencia Artificial (IA) en la evaluación del aprendizaje de idiomas está marcando un punto de inflexión en la pedagogía de segundas lenguas. Con herramientas como pruebas adaptativas, retroalimentación instantánea y evaluación de competencias comunicativas, la IA no solo está enriqueciendo la experiencia educativa, sino que también está redefiniendo la eficacia de las evaluaciones. Esto se evidencia en la percepción de los estudiantes: un 87% afirma que estas herramientas basadas en IA ofrecen una experiencia de aprendizaje más integral y enriquecedora, destacando su capacidad para adaptarse y personalizar el proceso educativo de manera inédita.

No obstante, es crucial recordar que la tecnología, incluida la IA, es una herramienta en apoyo del aprendizaje y no una solución en sí misma. Debe ser un complemento que asista a los educadores, no un dictador del proceso educativo.

En este contexto, el futuro de la evaluación en la enseñanza de idiomas se vislumbra como una integración efectiva de enfoques pedagógicos centrados en el aprendizaje con tecnologías de IA. Las evaluaciones en línea basadas en escenarios auténticos, potenciadas por IA, emergen como herramientas potentes, no solo para medir conocimiento, sino también como medios cruciales para fomentar el aprendizaje y el desarrollo de habilidades esenciales. Sin embargo, es esencial continuar explorando y perfeccionando la interacción con avatares y otros componentes de IA para garantizar una experiencia educativa completa y enriquecedora para todos los estudiantes.

Mirando hacia el futuro, es imprescindible seguir investigando sobre cómo estas tecnologías pueden adaptarse aún mejor a las necesidades individuales de los estudiantes y cómo pueden integrarse de manera más holística en el proceso educativo, asegurando que la tecnología sirva efectivamente a su propósito pedagógico y no solo como una novedad tecnológica.

7. BIBLIOGRAFÍA

Agarwal, S., Agarwal, B., & Gupta, R. (2022). Chatbots and virtual assistants: A bibliometric analysis. *Library Hi Tech*. https://doi.org/10.1108/LHT-09-2021-0330

Bardovi-Harlig, K. & Dornyei, Z. (1998) Do Language Learners Recognize Pragmatic Violations? Pragmatic versus Grammatical Awareness in Instructed L2 Learning. *TESOL Quarterly*, 32, 233-262.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5, 7-74.

Bloom, B. S. (1969). *Taxonomy of educational objectives: The classification of educational goals*. New York: Longman.

Bygate, M., & Samuda, V. (2009). Creating pressure in task pedagogy: The joint roles of field, purpose, and engagement within the interaction approach. In Mackey, A. and Polio, C., eds., Multiple Perspectives on Interaction: Second Language Research in Honor of Susan M. Gass (pp. 90–116). New York, NY: Routledge.

Carless, D. (2003). Learning-oriented assessment. Paper presented at the Evaluation and Assessment Conference, University of South Australia, Adelaide, November 25.

Dochy, F., Segers, M. S. R., & Dierick, S. (2002). Nuevas vías de aprendizaje y enseñanza y sus consecuencias: una nueva era de evaluación. *REDU Boletín*, 2(2), 13-30.

Merino, A (2022). Promoting Learning through Formative Performance Assessment and Feedback. Symposium on Language Pedagogy in Higher Education. University of Illinois.

Nelson, T. and Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125–173.

Richards, J. C., & Lockhart, C. (1994). Reflective Teaching in Second Language Classrooms. Cambridge University Press.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), Perspectives of curriculum evaluation (pp. 39-83). Chicago: Rand McNally.

Troyan, F., Adair-Hauck, B, & E. Glisan (2023) The integrated performance assessment: Twenty years and counting. ACTFL. Language Connects.