Evolución de las correcciones automatizadas en ELE: Análisis de Claude Sonnet 3.7 y 4.0 como evaluadores

Evolution of automated feedback in SFL: Analysis of Claude Sonnet 3.7 and 4.0 as evaluators

Evolução das correções automatizadas em ELE: Análise de Claude Sonnet 3.7 e 4.0 como avaliadores

Antoni Brosa-Rodríguez
Universitat Rovira i Virgili, España
antoni.brosa@urv.cat

Recibido: 28/06/2025
Aceptado: 19/12/2025
Publicado: 30/12/2025

Resumen: Esta investigación analiza la evolución de las capacidades de retroalimentación entre la inteligencia artificial específica para diferentes modelos de lenguaje Claude Sonnet en sus versiones 3.7 y 4.0 como herramientas de corrección para textos de estudiantes de español como lengua extranjera. Mediante análisis comparativo cualitativo de 15 textos del corpus CEDEL2, el estudio evalúa diferentes ítems: precisión en detección de errores, claridad explicativa, adecuación pedagógica y problemas detectados. Claude 4.0 incrementa la detección de errores en 17% (189 vs 161) y desarrolla mayor sofisticación en adaptación por niveles, concentrándose en errores fundamentales para principiantes mientras proporciona análisis exhaustivos para estudiantes avanzados. La versión más reciente introduce mejoras en organización estructural mediante formato tripartito "error → corrección → explicación". Sin embargo, presenta retrocesos pedagógicos preocupantes: elimina actividades complementarias características de Claude 3.7, degrada la retroalimentación motivacional a comentarios genéricos en tercera persona, y mantiene sesgos hacia variedades peninsulares e hipercorrección. Más problemáticas resultan las interferencias interlingüísticas que generan propuestas en español e inglés, generando un *spanglish* inadecuado. El análisis confirma que ninguna versión puede funcionar autónomamente sin mediación docente, estableciendo su rol óptimo como herramientas complementarias con supervisión pedagógica activa. Los hallazgos evidencian que la evolución tecnológica en inteligencia artificial educativa no constituye mejora lineal, revelando intercambios complejos entre sofisticación técnica y adecuación pedagógica.
Palabras clave: Enseñanza de idiomas; Tecnología educativa; Inteligencia artificial; Aprendizaje asistido por computadora

Abstract: This research analyses the evolution of feedback capabilities between specific artificial intelligence for different Claude Sonnet language models in versions 3.7 and 4.0 as correction tools for texts written by students of Spanish as a foreign language. Through a qualitative comparative analysis of 15 texts from the CEDEL2 corpus, the study evaluates different items: accuracy in error detection, explanatory clarity, pedagogical appropriateness, and problems

detected. Claude 4.0 increases error detection by 17% (189 vs 161) and develops greater sophistication in level adaptation, focusing on fundamental errors for beginners while providing comprehensive analysis for advanced students. The latest version introduces improvements in structural organisation through a tripartite format: 'error → correction → explanation'. However, it presents worrying pedagogical setbacks: it eliminates complementary activities characteristic of Claude 3.7, degrades motivational feedback to generic third-person comments, and maintains biases towards peninsular varieties and hypercorrection. More problematic are the interlinguistic interferences generated by proposals in Spanish and English, resulting in inappropriate Spanglish. The analysis confirms that neither version can function autonomously without teacher mediation, establishing their optimal role as complementary tools with active pedagogical supervision. The findings show that technological evolution in educational AI does not constitute linear improvement, revealing complex exchanges between technical sophistication and pedagogical adequacy.
Keywords: Language teaching; Educational technology; Artificial intelligence; Computer-assisted learning.

Resumo: Esta investigação analisa a evolução das capacidades de retroalimentação entre a inteligência artificial específica para diferentes modelos de linguagem Claude Sonnet nas suas versões 3.7 e 4.0 como ferramentas de correção para textos de estudantes de espanhol como língua estrangeira. Por meio de uma análise comparativa qualitativa de 15 textos do corpus CEDEL2, o estudo avalia diferentes itens: precisão na deteção de erros, clareza explicativa, adequação pedagógica e problemas detetados. O Claude 4.0 aumenta a deteção de erros em 17% (189 vs 161) e desenvolve maior sofisticação na adaptação por níveis, concentrando-se em erros fundamentais para iniciantes, ao mesmo tempo que fornece análises exaustivas para estudantes avançados. A versão mais recente introduz melhorias na organização estrutural por meio do formato tripartido "erro → correção → explicação". No entanto, apresenta retrocessos pedagógicos preocupantes: elimina atividades complementares características do Claude 3.7, degrada o feedback motivacional a comentários genéricos na terceira pessoa e mantém vieses em relação às variedades peninsulares e à hipercorreção. Mais problemáticas são as interferências interlinguísticas que geram propostas em espanhol e inglês, criando um spanglish inadequado. A análise confirma que nenhuma versão pode funcionar de forma autónoma sem a mediação do professor, estabelecendo o seu papel ideal como ferramentas complementares com supervisão pedagógica ativa. As conclusões evidenciam que a evolução tecnológica na IA educativa não constitui uma melhoria linear, revelando interações complexas entre sofisticação técnica e adequação pedagógica.
Palavras-chave: Ensino de idiomas; Tecnologia educativa; Inteligência artificial; Aprendizagem assistida por computador.

1. INTRODUCCIÓN

1.1. Planteamiento y justificación de la investigación

La corrección y retroalimentación de textos escritos constituyen una de las tareas fundamentales y más exigentes en la enseñanza del español como lengua extranjera (en adelante ELE). Esta actividad no solo requiere una inversión temporal considerable por parte del profesorado, sino que necesita un nivel de personalización y adaptación que resulta difícil de mantener en contextos educativos contemporáneos caracterizados por

el incremento de las ratios estudiante-profesor y la diversificación de las responsabilidades docentes (Buyse, 2014; Fernández, 2017).

En este escenario, la aparición de modelos de inteligencia artificial avanzados ha abierto nuevas posibilidades para el apoyo automatizado al proceso de corrección. Los desarrollos recientes en Large Language Models (LLM) han alcanzado niveles de sofisticación que permiten análisis textuales complejos y la generación de retroalimentación contextualizada y pedagógicamente orientada (Kasneci et al., 2023). Entre estos desarrollos, ChatGPT y Claude han emergido como una herramienta particularmente prometedora debido a su capacidad demostrada para manejar textos en español y proporcionar análisis lingüísticos detallados (García, 2024).

Sin embargo, la rápida evolución de estas tecnologías plantea cuestiones cruciales sobre la dirección y naturaleza de su desarrollo. ¿Representan las nuevas versiones mejoras inequívocas en términos pedagógicos? ¿Cómo evolucionan las capacidades específicas de detección, explicación y adaptación educativa? ¿Qué nuevas limitaciones o fortalezas emergen en el proceso de actualización tecnológica?

La necesidad de responder a estas preguntas se intensifica considerando que los educadores deben tomar decisiones informadas sobre qué herramientas implementar en sus contextos específicos. La asunción de que las versiones más recientes son automáticamente superiores requiere verificación empírica, especialmente en dominios especializados como la enseñanza de lenguas extranjeras donde las consideraciones pedagógicas pueden no alinearse directamente con los avances técnicos generales.

1.2. Marco teórico: retroalimentación correctiva en la enseñanza del ELE

La retroalimentación correctiva ha sido objeto de investigación intensiva en el campo de la adquisición de segundas lenguas, estableciéndose como un componente esencial del proceso de aprendizaje (Hattie y Timperley, 2007; Fernández, 2017; Bailini, 2020b; Mizumoto et al., 2024). En el contexto específico del español como lengua extranjera, la evolución teórica ha transitado desde enfoques centrados en la corrección explícita y exhaustiva hacia perspectivas más matizadas que consideran el error como parte natural del proceso de interlengua.

Bailini (2020a) propone una escala de regulación de estrategias de retroalimentación que la conceptualiza de forma efectiva como una integración de múltiples dimensiones: precisión en la detección, claridad explicativa, adecuación pedagógica y equilibrio entre corrección y motivación. Esta perspectiva multidimensional resulta particularmente relevante para evaluar herramientas automatizadas, ya que diferentes aspectos pueden evolucionar de manera independiente o incluso contradictoria.

La investigación contemporánea ha identificado principios fundamentales para la retroalimentación efectiva en ELE. La selectividad en la corrección emerge como un principio crucial: la corrección exhaustiva puede sobrecargar cognitivamente a los estudiantes, especialmente en niveles iniciales, mientras que la corrección selectiva y focalizada facilita el procesamiento y la incorporación de las correcciones (Bailini, 2020b). Asimismo, la adaptación al nivel del estudiante constituye otro principio

fundamental, requiriendo que las correcciones y explicaciones se ajusten tanto al estadio de competencia lingüística como al contexto de aprendizaje específico.

La dimensión afectiva de la retroalimentación ha ganado reconocimiento creciente en la literatura especializada. El equilibrio entre corrección y refuerzo positivo no constituye un aspecto cosmético del proceso, sino un factor determinante para mantener la motivación estudiantil y construir confianza en el proceso de aprendizaje, como así lo demuestran García Pujals y Lasagabaster (2019).

1.3. Inteligencia artificial y automatización de la retroalimentación lingüística

Los sistemas tutoriales inteligentes han constituido durante décadas un área de investigación activa en la intersección entre tecnología educativa y enseñanza de lenguas. Los primeros sistemas automatizados de corrección se caracterizaron por su efectividad en errores básicos y sistemáticos, pero mostraron limitaciones significativas en aspectos contextuales, pragmáticos y culturales (Cotterall, 2008; Ferreira y Kotz, 2010; Ranalli et al., 2017; Crossley et al., 2019).

El desarrollo de LLM ha revolucionado las posibilidades de automatización de la retroalimentación lingüística. Características como la comprensión contextual, la capacidad de adaptación, el multilingüismo y la generación de explicaciones contextualizadas han abierto perspectivas antes impensables para la personalización del apoyo educativo (Kasneci et al., 2023). Sin embargo, como señalan Coyne et al. (2023), estas capacidades emergentes requieren evaluación crítica específica para contextos educativos, ya que la sofisticación técnica no garantiza automáticamente la adecuación pedagógica.

Los estudios iniciales sobre la aplicación de modelos como ChatGPT en la enseñanza de lenguas han revelado tanto potencial significativo como limitaciones importantes, como puede ser la dificultad de establecer vínculos con el alumnado. Por otro lado, las Inteligencias Artificiales ayudan y facilitan la labor docente y estudiantil, como señala López Mata (2023), quien identifica capacidades prometedoras para la preparación de exámenes específicos como el DELE, mientras que Xiao y Zhi (2023) documentan percepciones estudiantiles mayoritariamente positivas hacia el uso de estas herramientas. Sin embargo, Ranalli (2021) advierte sobre problemas de confianza y precisión que pueden afectar la efectividad de la retroalimentación automatizada.

La evolución rápida de estos modelos plantea desafíos particulares para la investigación educativa (Arnold, 2000). Mientras que los estudios tradicionales pueden asumir estabilidad relativa en las herramientas evaluadas, los modelos de inteligencia artificial contemporáneos experimentan actualizaciones frecuentes que pueden alterar sustancialmente sus capacidades. Esta dinámica requiere enfoques de evaluación que puedan capturar tanto el estado actual como las tendencias evolutivas de estas tecnologías.

1.4. Objetivos y justificación del estudio

A pesar del interés creciente en la aplicación de modelos de inteligencia artificial a la enseñanza de ELE, al tratarse de un tema novedoso y actual, existen áreas no

abordadas, como la presente. Primero, existe una escasez de estudios que evalúen específicamente la evolución de las capacidades pedagógicas entre versiones consecutivas del mismo modelo. La mayoría de la investigación disponible se concentra en evaluaciones puntuales o comparaciones entre modelos diferentes, perdiendo la oportunidad de comprender las trayectorias de desarrollo tecnológico.

Segundo, la evaluación de la precisión lingüística de estos modelos en español, particularmente con textos producidos por estudiantes de ELE, permanece limitada. Los estudios existentes tienden a utilizar textos artificiales o se concentran en otras lenguas, dejando sin explorar comportamientos específicos relevantes para la enseñanza del español.

Tercero, la literatura carece de análisis detallados sobre la adecuación pedagógica de las explicaciones generadas por estos sistemas. Mientras que la capacidad de detección de errores ha recibido atención considerable, la calidad educativa de las explicaciones y su adaptación a diferentes niveles de competencia ha sido menos explorada sistemáticamente.

Finalmente, existe una ausencia notable de propuestas metodológicas concretas para la integración efectiva de estas herramientas en la práctica docente real. La mayor parte de la investigación disponible se mantiene en el nivel de evaluación de capacidades técnicas, sin traducir los hallazgos en recomendaciones prácticas para educadores.

Este estudio aborda estas lagunas mediante un análisis comparativo cualitativo y cuantitativo detallado entre Claude 3.7 y Claude 4.0, evaluando su evolución en cuatro dimensiones críticas: precisión en la detección de errores, claridad de las explicaciones, adecuación pedagógica y problemas detectados. Utilizando una muestra de textos auténticos del corpus CEDEL2, el estudio busca comprender no solo qué capacidades poseen estas herramientas, sino cómo evolucionan estas capacidades y qué implicaciones tiene esta evolución para la práctica educativa.

Los objetivos específicos del estudio incluyen: (1) evaluar comparativamente la precisión de detección de errores entre Claude 3.7 y 4.0 en textos de estudiantes de ELE; (2) analizar la evolución en la calidad y claridad de las explicaciones proporcionadas; (3) examinar los cambios en la adecuación pedagógica de la retroalimentación generada; (4) identificar nuevas fortalezas y limitaciones emergentes en la versión más reciente; y (5) formular recomendaciones específicas para la implementación educativa basadas en los hallazgos.

La relevancia de este estudio se fundamenta en la necesidad urgente de orientación empírica para educadores que deben navegar el paisaje cambiante de las herramientas de inteligencia artificial educativa. En un contexto donde las actualizaciones tecnológicas ocurren con frecuencia acelerada, la comprensión de las trayectorias de desarrollo resulta crucial para la toma de decisiones informadas sobre adopción e implementación tecnológica.

Además, este estudio contribuye al cuerpo emergente de investigación sobre la aplicación de inteligencia artificial en la enseñanza del ELE, un área que requiere urgentemente marcos teóricos y evidencia empírica específica. Los hallazgos pueden

informar tanto el desarrollo futuro de estas tecnologías como las estrategias pedagógicas para su integración efectiva.

A continuación, se presenta la metodología empleada para abordar estos objetivos, seguida por el análisis detallado de los resultados y su discusión en el contexto de las implicaciones pedagógicas más amplias. El estudio concluye con recomendaciones específicas para la implementación práctica y direcciones para la investigación futura en esta área de desarrollo acelerado.

2. METODOLOGÍA

Esta investigación adopta un enfoque de análisis comparativo para evaluar la efectividad de dos versiones consecutivas del modelo de inteligencia artificial Claude (3.7 y 4.0) como herramientas de retroalimentación para textos producidos por estudiantes de español como lengua extranjera. El diseño metodológico se estructura en torno a la evaluación sistemática de las capacidades correctivas de ambas versiones, permitiendo identificar tanto las mejoras como las limitaciones que caracterizan la evolución entre estas tecnologías.

2.1. Selección del corpus

Para garantizar la comparabilidad y representatividad de los datos, se analizaron 15 textos producidos por estudiantes de español como lengua extranjera, seleccionados según criterios específicos que aseguraran la diversidad y relevancia de la muestra:

1. Distribución por niveles de competencia: Los textos se distribuyeron equitativamente en tres grupos de cinco textos cada uno, correspondientes a nivel bajo (A1-A2), nivel intermedio (B1-B2) y nivel avanzado (C1-C2) según el Marco Común Europeo de Referencia para las Lenguas. Esta estratificación permitió evaluar la capacidad de ambas versiones de Claude para adaptar su retroalimentación a diferentes estadios de competencia lingüística.
2. Diversidad de lenguas maternas: La muestra incluyó textos de estudiantes con cinco lenguas maternas diferentes (árabe, inglés, japonés, alemán y neerlandés), asegurando la representación de diferentes familias lingüísticas y tipos de interferencia potencial. Esta diversidad resulta crucial para evaluar la capacidad de los modelos para manejar errores derivados de diferentes sistemas lingüísticos de partida.
3. Homogeneidad tipológica: Todos los textos correspondieron a producciones narrativas y descriptivas basadas en la descripción de una escena cinematográfica específica (fragmento de una película de Charlie Chaplin). Esta unificación tipológica eliminó variables relacionadas con el género textual, concentrando el análisis en las capacidades correctivas puras de los modelos.
4. Presencia de errores representativos: Los textos seleccionados contenían errores frecuentes en diferentes niveles lingüísticos (ortográfico, morfológico, sintáctico, léxico y pragmático), proporcionando un espectro

amplio de fenómenos para evaluar las capacidades de detección y corrección de ambas versiones de Claude.

2.2. Modelos de inteligencia artificial analizados

El estudio se centró en la evaluación comparativa de dos versiones consecutivas del modelo Claude Sonnet desarrollado por Anthropic:

Claude 3.7: Versión disponible en el momento inicial del estudio, representando el estado de desarrollo anterior del modelo. Esta versión constituye la línea base para la comparación evolutiva.

Claude 4.0: Versión más reciente disponible durante la fase de recolección de datos, representando los avances más actualizados en las capacidades del modelo. La selección de esta versión permite evaluar la dirección y naturaleza de la evolución tecnológica.

Ambas versiones fueron seleccionadas por representar el estado del arte en modelos de lenguaje conversacional y por su capacidad demostrada para procesar y generar texto en español con niveles elevados de competencia lingüística. La comparación entre versiones consecutivas del mismo modelo elimina variables relacionadas con arquitecturas fundamentalmente diferentes, concentrando el análisis en los cambios específicos introducidos en el proceso de desarrollo.

2.3. Protocolo de evaluación

Para asegurar la consistencia y comparabilidad de los resultados, se estableció un protocolo sistemático de evaluación que minimizara las variables externas y maximizara la validez de la comparación:

1. *Prompt*[1] estandarizado: Se diseñó un *prompt* idéntico para ambas versiones, solicitando análisis de errores con propósitos pedagógicos y didácticos. El texto exacto del *prompt* utilizado fue:

   > «Te adjunto un texto escrito por un estudiante que está aprendiendo español como lengua extranjera. Tenía que describir lo que veía en un vídeo que era un fragmento de Chaplin. Necesito que analices el texto para ver qué errores ha cometido. Debe servir para propósitos pedagógicos y didácticos, por tanto, estaría bien que hubiera explicación, propuesta de corrección, etc. Debes hacer lo que haría un profesor. Tienes que tener en cuenta el nivel de los estudiantes y su contexto y punto de partida lingüístico. Envía la versión final para darles a ellos.»

Este *prompt* fue cuidadosamente diseñado para solicitar comportamiento pedagógico específico, incluyendo detección de errores, explicaciones conceptuales, propuestas de corrección y adaptación al nivel del estudiante.

[1] Se entiende por *prompt* una instrucción, pregunta o solicitud que se le da a un sistema de inteligencia artificial para que genere una respuesta o realice una tarea específica y para obtener resultados relevantes y precisos. La calidad de la respuesta depende en gran medida de la claridad y detalle de dicho *prompt*.

2. Procedimiento de aplicación: Cada uno de los 15 textos fue analizado independientemente por ambas versiones de Claude, generando un total de 30 análisis (15 textos × 2 modelos). Los análisis se realizaron en sesiones separadas para evitar cualquier posible influencia entre las evaluaciones de las diferentes versiones.
3. Control de variables: Se mantuvieron constantes todas las condiciones de evaluación excepto la versión del modelo utilizada, incluyendo el horario de aplicación, el formato de presentación de los textos, y la formulación exacta de las instrucciones.

2.4. Criterios de análisis

El análisis cualitativo de las retroalimentaciones generadas se estructuró en torno a cuatro criterios fundamentales, desarrollados a partir de la literatura especializada en retroalimentación efectiva para la enseñanza de lenguas extranjeras (Bailini, 2020a; Fernández, 2017):

1. Precisión en la detección de errores: Este criterio evaluó la capacidad de cada versión para identificar correctamente los errores presentes en los textos, considerando tanto la sensibilidad (capacidad de detectar errores reales) como la especificidad (evitar falsos positivos). Se analizaron también los patrones de detección según el tipo de error y el nivel del estudiante.
2. Claridad de las explicaciones: Se evaluó la calidad de las explicaciones proporcionadas, incluyendo la precisión conceptual, la contextualización de las correcciones, la adecuación de la ejemplificación, y la coherencia en la presentación de la información. Este criterio resulta crucial para que la retroalimentación cumpla efectivamente su función pedagógica.
3. Adecuación pedagógica: Se analizó la adaptación de la retroalimentación al nivel del estudiante, la progresión apropiada de la dificultad, el equilibrio entre corrección y refuerzo positivo, y la inclusión de elementos motivacionales. Este criterio evalúa la comprensión de los modelos sobre los principios fundamentales del proceso de enseñanza-aprendizaje.
4. Problemas detectados: Se identificaron sistemáticamente las dificultades, limitaciones o efectos adversos introducidos por cada versión, incluyendo hipercorrección, sesgos dialectales, inconsistencias, y cualquier otro aspecto que pudiera interferir negativamente con el proceso de aprendizaje.

Para cada criterio se realizó un análisis detallado que identificó patrones recurrentes, ejemplos representativos, y diferencias significativas entre las versiones. Este enfoque cualitativo permitió una comprensión profunda de las capacidades y limitaciones de cada modelo, más allá de las métricas puramente cuantitativas.

2.5. Procedimientos del estudio

El análisis de los datos se llevó a cabo mediante un proceso sistemático de revisión cualitativa que combinó análisis de contenido temático con comparación constante entre las versiones evaluadas. Cada retroalimentación fue analizada

independientemente según los cuatro criterios establecidos, identificando tanto fortalezas como debilidades específicas. Este tipo de análisis permite ahondar en la calidad de las retroalimentaciones generadas y poder detectar invariancias en el tipo de respuestas.

Posteriormente se realizó un análisis comparativo directo entre las respuestas de ambas versiones para cada texto, permitiendo identificar diferencias en el comportamiento correctivo, patrones de evolución, y cambios en el enfoque pedagógico. Este análisis comparativo resultó esencial para comprender la naturaleza de la evolución entre versiones y sus implicaciones para la aplicación educativa.

Los resultados cuantitativos se han presentado en cantidad dentro de los distintos tipos de errores detectados, relativos a diferentes cuestiones: puntuación, acentos, géneros, verbos, vocabulario, estructura oracional, ortografía y preposiciones. Se utilizaron como datos complementarios para apoyar y contextualizar los hallazgos cualitativos, pero el énfasis metodológico se mantuvo en la comprensión profunda de las capacidades pedagógicas de cada versión.

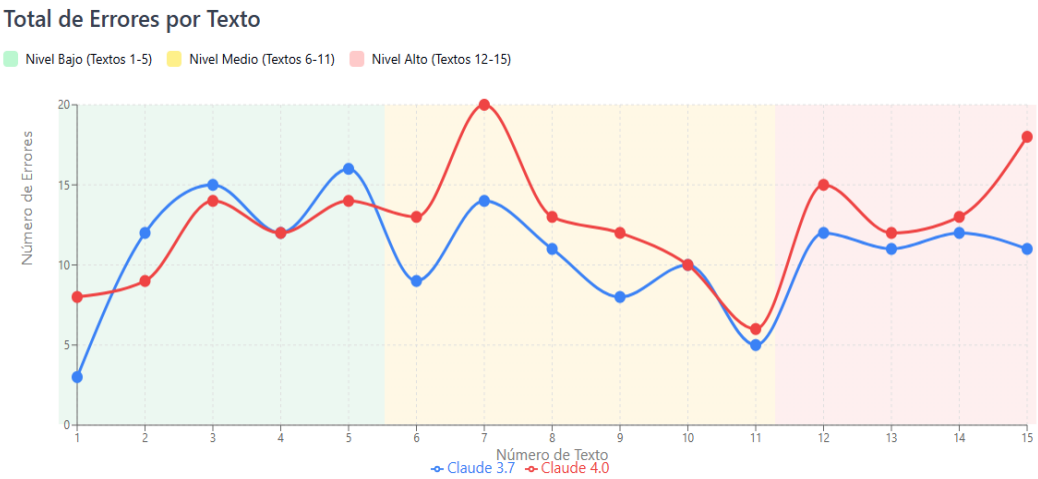
Esta aproximación metodológica permite una evaluación integral de las herramientas analizadas, proporcionando tanto una caracterización detallada de sus capacidades como una base sólida para formular recomendaciones sobre su implementación efectiva en contextos educativos reales.

3.RESULTADOS

El análisis comparativo entre Claude 3.7 y Claude 4.0 como herramientas de corrección para estudiantes de español como lengua extranjera reveló patrones significativos tanto en términos cuantitativos como cualitativos. Los datos obtenidos mostraron diferencias sustanciales en el comportamiento de ambas versiones, evidenciando una evolución compleja que combina mejoras técnicas con nuevos desafíos pedagógicos.

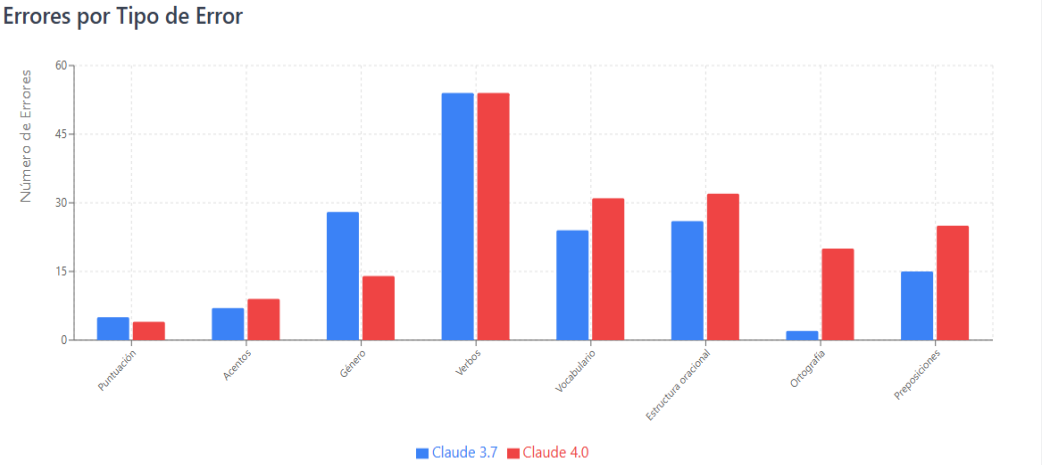
Desde una perspectiva cuantitativa —Figura 1—, Claude 4.0 detectó 189 errores frente a los 161 identificados por Claude 3.7, representando un incremento moderado del 17% en la detección total. Sin embargo, este incremento no se distribuyó uniformemente a través de todos los niveles de competencia. Como se observa en la Figura 1, Claude 4.0 mostró un comportamiento diferenciado: detectó proporcionalmente menos errores en textos de nivel inicial (A1-A2) mientras incrementó significativamente la identificación en niveles intermedios y avanzados (B1-C2). Este patrón contrastó con Claude 3.7, que mantuvo una distribución más homogénea de correcciones independientemente del nivel del estudiante.

Figura 1
Total de errores detectados en Claude 3.7 y 4.0 por texto y nivel



Fuente: Elaboración propia

Figura 2
Errores detectados por Claude 3.7 y 4.0 clasificados por tipo de error



Fuente: Elaboración propia

La Figura 2 reveló disparidades notables en la detección por categorías de error. Claude 4.0 superó consistentemente a su versión anterior en todas las categorías analizadas, con diferencias especialmente marcadas en aspectos estructurales como verbos (55 vs 27 errores), estructura oracional (32 vs 25 errores) y preposiciones (24 vs 15 errores). Resultó particularmente llamativo que Claude 4.0 detectara menos errores de concordancia de género que Claude 3.7 (3 vs 8 errores), siendo la única categoría donde la versión más reciente identificó menos problemas.

3.1. Precisión en la detección de errores

3.1.1. Claude 3.7

Claude 3.7 demostró una precisión general elevada en la identificación de errores básicos, con particular efectividad en la detección de problemas ortográficos, incluyendo acentuación («encontro → encontró», «despues → después»), errores morfosintácticos como problemas de concordancia y uso incorrecto de preposiciones («vio a una carta → vio una carta"), y errores de expresión que afectaban la adecuación comunicativa.

No obstante, el análisis reveló limitaciones significativas en su comportamiento correctivo. Claude 3.7 mostró tendencias de sobredetección, señalando en textos de niveles iniciales errores relacionados con estructuras gramaticales avanzadas que excedían las expectativas para ese estadio de aprendizaje. La caracterización imprecisa de errores constituyó otro problema recurrente, como cuando clasificó «Un bebe → Un bebé» como error de concordancia de género en lugar de error de acentuación. Adicionalmente, se observaron casos de hipercorrección donde se marcaron como erróneas construcciones gramaticalmente aceptables, y un sesgo hacia variedades peninsulares que llevó a corregir usos válidos en otras variantes del español («ella se enoja → ella se enfada»).

3.1.2. Claude 4.0

Claude 4.0 mantuvo la alta precisión de su versión anterior en la detección de errores básicos, pero introdujo mejoras y nuevas problemáticas. Entre las fortalezas más destacadas se encontró su capacidad de ajusta la corrección al nivel de competencia inferido del estudiante. Esta versión se concentró en errores fundamentales en niveles iniciales mientras proporcionó análisis más exhaustivos en niveles avanzados, sugiriendo una comprensión mejorada del proceso pedagógico.

La precisión en errores estructurales oracionales representó otra mejora notable, con incrementos significativos en la detección de problemas verbales, estructura oracional y uso de preposiciones. Claude 4.0 tendió a realizar correcciones integrales, abordando múltiples aspectos problemáticos en una sola intervención, lo que explicó la reducción en la detección de errores de concordancia de género: al corregir problemas estructurales mayores o sugerir cambios léxicos, resolvió automáticamente los problemas de concordancia sin necesidad de señalarlos específicamente.

Sin embargo, persistieron limitaciones importantes heredadas de la versión anterior. La hipercorrección de variedades diatópicas continuó siendo problemática, marcando como erróneos términos válidos en variedades no peninsulares como «carriola» (mexicanismo) o «agarra», perpetuando la tradición discriminatoria a un sesgo dialectal no justificado. Claude 4.0 siguió realizando correcciones prematuras, señalando construcciones con subjuntivo en textos de nivel A1 («amo → ame»), y mantuvo generalidades presentes en la versión 3.7 como el error persistente de corregir «negro y blanco» por «blanco y negro» sin aportar la información de uso y extensión específica de esta construcción, algo que podría causar confusión o errores futuros.

Más preocupante resultó la aparición de nuevos tipos de problemáticas, como interferencias lingüísticas evidenciadas en casos donde se detectaron propuestas de

spanglish, como la corrección de «infintil» sugiriendo el uso de «child», revelando cruces inapropiados entre lenguas.

3.2. Claridad de las explicaciones

3.2.1. Claude 3.7

En términos de calidad explicativa, Claude 3.7 ofreció retroalimentación contextualizada con refuerzo positivo, destacando sistemáticamente aspectos positivos de los textos analizados, aunque ocasionalmente careció de ejemplificación concreta. Su formato estructurado en listas facilitó la identificación de diferentes elementos, pero a veces resultó en excesiva compartimentación que dificultó la percepción de problemas interrelacionados.

Las explicaciones gramaticales tendieron a ser breves y no siempre consistentes. Mientras que en algunos casos proporcionó explicaciones gramaticales concisas, en otros se limitó a señalar errores sin justificar las correcciones propuestas. Particularmente problemático fue su enfoque memorístico para ciertos aspectos como las reglas de acentuación, donde no explicó los principios subyacentes.

3.2.2. Claude 4.0

Claude 4.0 introdujo mejoras sustanciales en la organización y claridad estructural de sus explicaciones. Implementó consistentemente una estructura tripartita clara de «error → corrección → explicación» que facilitó significativamente la comprensión del proceso correctivo. Esta sistematización mejorada presentó las correcciones en formato de lista organizada que permitió a los estudiantes seguir un orden lógico de revisión, representando una clara evolución respecto a la versión anterior.

Las explicaciones se volvieron más contextualizadas, proporcionando justificaciones gramaticales más precisas para cada corrección propuesta. Sin embargo, esta mejora en la organización formal introdujo nuevas problemáticas. El etiquetado erróneo de categorías constituyó un problema significativo: los epígrafes de las secciones no siempre correspondieron con el tipo de error analizado, como, por ejemplo, cuando clasificó un error de acentuación («encontró» sin acento) bajo la categoría «concordancia y cohesión», generando confusión conceptual que podía interferir con el aprendizaje.

Adicionalmente, se observó inconsistencias en el nivel de detalle proporcionado para errores similares, sin criterios aparentes que justificaran estas diferencias, lo que podría generar incertidumbre en los estudiantes sobre la importancia relativa de diferentes tipos de errores.

3.3. Adecuación pedagógica

3.3.1. Claude 3.7

La evaluación de la adecuación pedagógica de Claude 3.7 reveló tanto fortalezas como limitaciones importantes. Entre los aspectos positivos destacó su capacidad general de adaptación al nivel presumido del estudiante, aunque con las excepciones ya mencionadas respecto a la sobredetección. La retroalimentación incorporó sistemáticamente elementos motivadores y reconocimiento de logros, contribuyendo potencialmente a un ambiente de aprendizaje positivo.

Un aspecto particularmente valioso fue la inclusión consistente de sugerencias para actividades adicionales adaptadas a los problemas detectados, cerrando cada retroalimentación con preguntas o ejercicios complementarios. Sin embargo, se observaron limitaciones como el uso excesivo de terminología técnica en niveles iniciales y ocasional descontextualización de las reformulaciones propuestas.

3.3.2. Claude 4.0

Claude 4.0 presentó una evolución compleja en términos de adecuación pedagógica, combinando mejoras significativas con retrocesos evidentes. La mejora más notable fue su capacidad de adaptación cuantitativa por nivel, demostrando mayor sofisticación al modular la cantidad de correcciones según el nivel inferido del estudiante, evitando sobrecargar a principiantes mientras proporcionó análisis detallados a estudiantes avanzados.

El formato sistemático de presentación constituyó otra mejora, ya que apoyó efectivamente el proceso de revisión y autocorrección, facilitando que los estudiantes navegaran por las correcciones de manera organizada.

Sin embargo, Claude 4.0 sacrificó elementos pedagógicos valiosos presentes en su versión anterior. La ausencia de actividades complementarias representó una pérdida significativa, eliminando oportunidades para reforzar los aspectos corregidos mediante ejercicios personalizados. Más problemática resultó la gestión inadecuada de la retroalimentación positiva: la información motivacional apareció relegada al final de la retroalimentación, fue extremadamente breve y se presentó en tercera persona, contrastando marcadamente con la personalización solicitada en las instrucciones del estudio.

La generalidad motivacional constituyó otro retroceso, con elementos de refuerzo positivo que carecieron de especificidad y personalización, perdiendo la efectividad motivacional que caracterizó a Claude 3.7.

3.4. Problemas detectados

3.4.1. Claude 3.7

El análisis de Claude 3.7 identificó varios problemas sistemáticos que afectaron la calidad de su retroalimentación. Las contradicciones internas representaron una limitación significativa, como cuando ofreció una versión corregida con reestructuración de oraciones y luego propuso como actividad complementaria realizar esa misma reestructuración.

La señalización ambigua constituyó otra problemática, señalando frases correctas sin clarificar por qué aparecían en la sección de errores, generando confusión potencial. El fenómeno de *horror vacui* se manifestó en la tendencia a buscar aspectos mejorables incluso en textos de nivel básico donde ciertas imperfecciones podrían considerarse admisibles para el estadio de aprendizaje.

3.4.2. Claude 4.0

Claude 4.0 heredó algunos problemas de su versión anterior mientras introdujo nuevas limitaciones. El *horror vacui* persistió, manteniendo la tendencia a identificar aspectos mejorables incluso cuando no fueron pedagógicamente necesarios. El sesgo dialectal continuó siendo problemático, privilegiando variedades peninsulares sobre otras igualmente válidas y perpetuando discriminación lingüística.

Los nuevos problemas identificados resultaron particularmente preocupantes. Las interferencias lingüísticas entre idiomas representaron una regresión significativa, con la aparición de sugerencias inapropiadas que mezclaron idiomas de manera inadecuada. El desequilibrio en priorización se manifestó en el tratamiento inconsistente de errores similares, sugiriendo falta de criterios estables de evaluación.

Finalmente, la reducción del componente motivacional constituyó un retroceso pedagógico notable, con una disminución significativa en la calidad y personalización de la retroalimentación positiva que había caracterizado favorablemente a Claude 3.7.

La evolución de Claude 3.7 a 4.0 reveló así un intercambio complejo entre precisión técnica y adecuación pedagógica integral, donde las mejoras en aspectos cuantitativos y organizacionales coexistieron con retrocesos en elementos cruciales para el proceso de aprendizaje como la motivación personalizada y las actividades complementarias.

4.DISCUSIÓN

Los resultados obtenidos revelan una evolución compleja en las capacidades de Claude como herramienta de retroalimentación en la enseñanza del español como lengua extranjera. La comparación entre las versiones 3.7 y 4.0 muestra un panorama donde las mejoras técnicas coexisten con retrocesos pedagógicos significativos, planteando cuestiones fundamentales sobre la dirección del desarrollo de estas tecnologías educativas.

4.1. Interpretación de resultados

El incremento del 17% en la detección total de errores por parte de Claude 4.0 (189 vs 161) podría interpretarse inicialmente como una mejora en la capacidad analítica del sistema. Sin embargo, el análisis cualitativo revela que esta mejora cuantitativa no se traduce necesariamente en mayor calidad pedagógica. La distribución diferenciada observada de correcciones según el nivel del estudiante en Claude 4.0 sugiere una comprensión más sofisticada del proceso de aprendizaje, alineándose con los principios de retroalimentación selectivo y adaptado propuestos por Bailini (2020b).

Esta capacidad de modular la cantidad de correcciones según el nivel inferido representa un avance significativo respecto a la tendencia de sobredetección generalizada observada en Claude 3.7.

La reducción en la detección de errores de concordancia de género en Claude 4.0 (3 vs 8 errores) ilustra un fenómeno particularmente interesante: la evolución hacia correcciones más integrales que abordan múltiples aspectos problemáticos simultáneamente. Este comportamiento sugiere que Claude 4.0 ha desarrollado una comprensión más holística de los problemas textuales, resolviendo automáticamente ciertos errores en el proceso de corregir aspectos estructurales más amplios. Esta aproximación, aunque técnicamente más sofisticada, puede privar a los estudiantes de oportunidades de aprendizaje específicas sobre aspectos gramaticales concretos. Por ejemplo, a la hora de señalar un error de ortografía quizá también cambia el léxico y las concordancias, lo que puede invisibilizar el problema ortográfico, al quedar relegado en una sustitución más grande.

El incremento notable en la detección de errores estructurales (verbos, estructura oracional, preposiciones) por parte de Claude 4.0 coincide con las observaciones de Coyne et al. (2023) sobre el potencial de los sistemas de inteligencia artificial para identificar patrones complejos en la producción lingüística. Sin embargo, la aparición de interferencias lingüísticas como la sugerencia de «child» para corregir «infintil» representa una limitación significativa que sugiere problemas en el entrenamiento o la arquitectura del modelo más reciente.

La persistencia del sesgo hacia variedades peninsulares en ambas versiones confirma las observaciones de Xiao y Zhi (2023) sobre los desequilibrios en los datos de entrenamiento de los modelos de lenguaje. La corrección sistemática de términos como «carriola» o «agarra» como si fueran errores perpetúa una perspectiva monocéntrica del español que contradice los principios de pluricentrismo lingüístico ampliamente aceptados en la didáctica contemporánea del español como lengua extranjera.

4.2. Implicaciones pedagógicas

La evolución observada entre Claude 3.7 y 4.0 plantea cuestiones fundamentales sobre la dirección del desarrollo tecnológico en el ámbito educativo. Mientras Claude 4.0 muestra mejoras evidentes en organización y sistematización de la retroalimentación, sacrifica elementos pedagógicos cruciales que caracterizaban positivamente a su versión anterior como, por ejemplo, explicaciones más directas y aisladas.

La eliminación de actividades complementarias en Claude 4.0 constituye una limitación relevante desde la perspectiva del aprendizaje autónomo. Como señalan Benson (2006) y García Pujals y Lasagabaster (2019), el desarrollo de la autonomía del estudiante constituye un componente esencial del aprendizaje efectivo de lenguas extranjeras. Las actividades complementarias proporcionadas por Claude 3.7 ofrecían oportunidades valiosas para que los estudiantes profundizaran en los aspectos corregidos, facilitando la transferencia del conocimiento y la consolidación del aprendizaje.

La degradación en la calidad de la retroalimentación motivacional observada en Claude 4.0 resulta particularmente llamativa. La retroalimentación positiva personalizada no constituye un elemento cosmético del proceso de corrección, sino un componente fundamental para mantener la motivación del estudiante y construir confianza en el proceso de aprendizaje. Como destaca Bailini (2020a), el equilibrio entre corrección y motivación resulta crucial para que la retroalimentación cumpla efectivamente su función pedagógica.

El formato tripartito «error → corrección → explicación» implementado consistentemente por Claude 4.0 representa, no obstante, una mejora significativa en la claridad estructural de la retroalimentación. Esta organización facilita que los estudiantes comprendan tanto qué está mal como por qué está mal, apoyando el desarrollo de la conciencia metalingüística. Sin embargo, los problemas de etiquetado categorial observados pueden generar confusión conceptual que interfiera con este proceso. Surge así una cuestión de fondo: si la optimización estructural lograda mediante la inteligencia artificial puede —o debe— prevalecer sobre el potencial formativo, motivacional y autónomo que caracteriza al acompañamiento humano.

4.3. Limitaciones tecnológicas y pedagógicas

Los resultados evidencian que, pese a las mejoras técnicas, ambas versiones de Claude mantienen limitaciones fundamentales que cuestionan su capacidad para funcionar autónomamente como herramientas de retroalimentación. El fenómeno de *horror vacui* identificado en ambas versiones contradice los principios de retroalimentación selectiva y puede sobrecargar cognitivamente a los estudiantes, especialmente en niveles iniciales o directamente favorecer que se inventen errores inexistentes.

Las interferencias lingüísticas detectadas en Claude 4.0 representan un retroceso técnico significativo que sugiere problemas en la gestión multilingüe del modelo. Estas interferencias no solo proporcionan retroalimentación incorrecta, sino que pueden inducir confusión profunda a los estudiantes sobre las normas de la lengua meta.

La inconsistencia en los criterios de corrección observada en ambas versiones plantea serias dudas sobre la confiabilidad de estos sistemas. La variabilidad en el tratamiento de errores similares sugiere que los modelos carecen de criterios estables de evaluación, lo que puede generar inseguridad en los estudiantes sobre qué construcciones son realmente apropiadas.

4.4. Hacia un modelo integrado de implementación para el uso de inteligencia artificial en la retroalimentación

Los hallazgos de este estudio refuerzan la necesidad de un enfoque integrado que reconozca tanto las capacidades como las limitaciones de estas tecnologías. Basándose en los resultados observados, se propone un modelo de implementación que maximice las fortalezas de cada versión mientras mitigue sus debilidades.

Fase 1: Selección estratégica de herramientas. El docente debe seleccionar la versión más apropiada según el contexto específico. Claude 3.7 resulta más eficaz para

estudiantes que requieren mayor apoyo motivacional y actividades complementarias, mientras que Claude 4.0 puede ser más efectivo para estudiantes avanzados que se beneficien de su análisis más detallado y estructura organizativa mejorada.

Fase 2: Mediación pedagógica activa. Los resultados confirman que ninguna versión puede funcionar autónomamente sin supervisión docente. El profesor debe revisar la retroalimentación generada, corrigiendo errores de categorización, eliminando correcciones inapropiadas de variedades diatópicas, y añadiendo contexto pedagógico cuando sea necesario.

Fase 3: Desarrollo de competencia crítica estudiantil. Los estudiantes deben desarrollar capacidades evaluativas para interactuar críticamente con la retroalimentación automatizada. Esto incluye comparar correcciones con otras fuentes, cuestionar sugerencias que parezcan inconsistentes, y mantener una perspectiva escéptica hacia las propuestas de cambio.

4.5. Perspectivas y recomendaciones para el desarrollo de la retroalimentación automatizada en ELE

La evolución observada de Claude 3.7 a 4.0 sugiere que el desarrollo futuro de estas tecnologías tendría que equilibrar mejor las mejoras técnicas con las necesidades pedagógicas. Las versiones futuras deberían:

- Mantener o mejorar las capacidades de motivación personalizada y generación de actividades complementarias.
- Desarrollar mayor sensibilidad hacia la variación dialectal del español.
- Implementar criterios más consistentes y transparentes de evaluación.
- Reducir las interferencias lingüísticas y mejorar la gestión multilingüe.

Adicionalmente, se necesita investigación longitudinal que evalúe el impacto real de estas herramientas en el aprendizaje estudiantil, siguiendo las líneas de trabajo iniciadas por Feng Teng (2024) y Slamet (2024). Solo mediante estudios empíricos que midan el aprendizaje efectivo será posible determinar si las mejoras técnicas observadas se traducen en beneficios pedagógicos reales.

La integración efectiva de la inteligencia artificial en la enseñanza del español como lengua extranjera requiere un enfoque crítico y reflexivo que reconozca tanto el potencial transformador como las limitaciones inherentes de estas tecnologías. El objetivo no puede ser reemplazar la mediación humana, sino crear sinergias que potencien las capacidades tanto de las herramientas tecnológicas como de los educadores, siempre en servicio del objetivo fundamental: facilitar el desarrollo de la competencia comunicativa de los estudiantes en un mundo cada vez más interconectado y multilingüe.

5.CONCLUSIONES

El análisis comparativo entre Claude 3.7 y Claude 4.0 como herramientas de corrección para estudiantes de español como lengua extranjera revela una evolución tecnológica compleja que no puede caracterizarse simplemente como una mejora lineal.

Los resultados obtenidos muestran que, si bien Claude 4.0 introduce avances significativos en ciertos aspectos técnicos y organizativos, también presenta retrocesos importantes en dimensiones pedagógicas fundamentales, evidenciando que el progreso en inteligencia artificial no garantiza automáticamente mejoras en la aplicabilidad educativa.

Desde una perspectiva cuantitativa, Claude 4.0 demuestra una capacidad mejorada de detección de errores, identificando un 17% más de problemas que su versión anterior (189 vs 161 errores). Sin embargo, la contribución más valiosa de esta versión radica en su comportamiento diferenciado según el nivel del estudiante: la concentración en errores fundamentales para principiantes y el análisis más exhaustivo para estudiantes avanzados sugiere una comprensión más sofisticada del proceso pedagógico. Esta capacidad de modulación representa un avance significativo hacia la personalización de la retroalimentación automatizada.

En términos de organización y claridad estructural, Claude 4.0 establece un nuevo estándar con su implementación consistente del formato tripartito «error → corrección → explicación». Esta sistematización facilita significativamente la navegación de la retroalimentación por parte de los estudiantes y apoya el desarrollo de la conciencia metalingüística. La estructura organizativa mejorada constituye una evolución clara respecto a Claude 3.7 y demuestra el potencial de estas tecnologías para presentar información compleja de manera accesible.

No obstante, estos avances técnicos coexisten con limitaciones preocupantes que cuestionan la dirección del desarrollo tecnológico. La eliminación de actividades complementarias en Claude 4.0 representa una pérdida pedagógica significativa que contradice los principios establecidos de aprendizaje autónomo y transferencia de conocimiento. Estas actividades no constituían elementos accesorios en Claude 3.7, sino componentes esenciales que facilitaban la consolidación del aprendizaje y el desarrollo de la autonomía estudiantil.

Más problemática resulta la degradación en la calidad de la retroalimentación motivacional observada en Claude 4.0. La retroalimentación positiva relegada al final, presentada de manera genérica y en tercera persona, contrasta marcadamente con las necesidades de personalización y motivación identificadas en la literatura especializada. Esta regresión sugiere que las mejoras en organización técnica se han logrado a expensas de la comprensión de los aspectos afectivos del aprendizaje.

Ambas versiones mantienen limitaciones fundamentales que persisten a través de las actualizaciones. El sesgo hacia variedades peninsulares del español, manifestado en la corrección sistemática de términos válidos como «carriola» o «agarra», perpetúa una perspectiva monocéntrica que contradice los principios contemporáneos de enseñanza pluricéntrica del español. Esta limitación no solo refleja desequilibrios en los datos de entrenamiento, sino que puede transmitir a los estudiantes una visión empobrecida de la riqueza dialectal del español.

El fenómeno de *horror vacui* (necesidad de señalar errores, existan o no) identificado en ambas versiones revela una comprensión limitada del proceso pedagógico. La tendencia a identificar aspectos mejorables incluso cuando no son

pedagógicamente apropiados para el nivel del estudiante puede sobrecargar cognitivamente a los aprendices y desviar la atención de errores verdaderamente prioritarios. Esta limitación sugiere que, pese a los avances técnicos, los modelos aún carecen de una comprensión profunda de los principios de retroalimentación selectiva y adaptada.

Particularmente preocupante en Claude 4.0 es la aparición de interferencias lingüísticas, que representa una regresión técnica significativa. Estas interferencias no solo proporcionan retroalimentaciones incorrectas, sino que pueden generar confusión fundamental sobre las normas de la lengua meta, sugiriendo problemas en la gestión multilingüe del modelo más reciente.

Los problemas de etiquetado categorial observados en Claude 4.0 ilustran cómo las mejoras en organización pueden introducir nuevos tipos de confusión conceptual. Esta inconsistencia entre categorización y contenido puede interferir con el desarrollo de la conciencia metalingüística que el formato estructurado pretende facilitar.

Los hallazgos confirman categóricamente que ninguna de las versiones analizadas puede funcionar autónomamente como sustituto de la retroalimentación docente. Las limitaciones identificadas —desde sesgos dialectales hasta interferencias lingüísticas e inconsistencias evaluativas— establecen claramente que el rol óptimo de estas tecnologías es el de herramientas complementarias que requieren mediación pedagógica activa. Esta conclusión refuerza la importancia del docente como mediador crítico que contextualiza, valida y, cuando es necesario, corrige la retroalimentación automatizada.

Sin embargo, reconocer estas limitaciones no debe oscurecer el potencial significativo que ambas versiones ofrecen para apoyar el proceso de enseñanza-aprendizaje. La capacidad de proporcionar retroalimentación inmediata y estructurada, disponible fuera del horario de clase, puede facilitar el desarrollo de la autonomía estudiantil y complementar efectivamente la atención docente. La clave radica en implementar estas herramientas con plena conciencia de sus capacidades y limitaciones.

El análisis comparativo sugiere que la evolución futura de estas tecnologías debería orientarse hacia un equilibrio más efectivo entre sofisticación técnica y sensibilidad pedagógica. Las versiones futuras deberían mantener o recuperar elementos valiosos como las actividades complementarias y la retroalimentación motivacional personalizado, mientras desarrollan mayor conciencia de la variación dialectal y criterios más consistentes de evaluación o deberá ser incluida esta de un modo más forzado en el *prompt.*

La integración efectiva de inteligencia artificial en la enseñanza del español como lengua extranjera requiere un enfoque crítico y reflexivo que reconozca que el progreso tecnológico no es sinónimo de mejora pedagógica. Los educadores deben mantener un papel activo como evaluadores críticos de estas tecnologías, seleccionando y adaptando las herramientas según las necesidades específicas de sus contextos educativos.

En el panorama actual, Claude 3.7 y Claude 4.0 representan estadios diferentes en la evolución de las herramientas de retroalimentación automatizada, cada uno con

fortalezas y debilidades específicas. Claude 3.7 destaca por su enfoque integral que incluye motivación y actividades complementarias, mientras que Claude 4.0 sobresale en organización y análisis diferenciado por niveles. La elección entre ambas versiones debe basarse en las necesidades específicas del contexto educativo y los objetivos pedagógicos particulares.

El futuro de la enseñanza del español como lengua extranjera evolucionará probablemente hacia modelos integrados donde la experiencia humana y las capacidades de la inteligencia artificial se complementen sinérgicamente. En este escenario, el desarrollo de la competencia crítica tanto de docentes como de estudiantes para evaluar y utilizar efectivamente estas tecnologías constituirá una habilidad fundamental. Solo mediante esta integración reflexiva y crítica podremos aprovechar el potencial transformador de la inteligencia artificial mientras preservamos los elementos esencialmente humanos que caracterizan la educación de calidad.

FINANCIACIÓN: Esta investigación no recibió ninguna financiación externa.

REFERENCIAS BIBLIOGRÁFICAS

Arnold, J. (2000). *La dimensión afectiva en el aprendizaje de idiomas.* Colección Cambridge de didáctica de lenguas. Edinumen.

Bailini, S. (2020a). El *feedback* como herramienta didáctica para el desarrollo de la autonomía en la adquisición de lenguas extranjeras. *Philologia Hispalensis, 34*, 25-39. https://dx.doi.org/10.12795/PH.2020.v34.i01.02

Bailini, S. (2020b). El *feedback* interactivo y la adquisición del español como lengua extranjera. Mimesis.

Benson, P. (2006). Autonomy in language teaching and learning. Language Teaching, 40, 21-40. https://doi.org/10.1017/S0261444806003958

Buyse, K. (2014). Una hoja de ruta para integrar las TIC en el desarrollo de la expresión escrita: recursos y resultados. *Journal of Spanish Language Teaching, 1*(1), 101-115. https://doi.org/10.1080/23247797.2014.898516

Coterall, S. (2008). Aprendientes de lenguas y autoevaluación. *marcoELE, 7*. http://marcoele.com/descargas/7/cotterall.pdf

Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M. e Inui, K. (2023). *Analyzing the performance of GPT-3.5 and GPT-4 in grammatical error correction*. arXiv:2303.14342. https://doi.org/10.48550/arXiv.2303.14342

Crossley, S. A., Bradfield, F. y Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research, 11*(2), 251-270. https://doi.org/10.17239/jowr-2019.11.02.01

Feng Teng, M. (2024). «ChatGPT is the companion, not enemies»: EFL learners' perceptions and experiences in using ChatGPT for *feedback* in writing. *Computers*

and Education: Artificial Intelligence, 7. https://doi.org/10.1016/j.caeai.2024.100270

Fernández, S. (2017). Evaluación y aprendizaje. *MarcoELE: Revista de Didáctica Español Lengua Extranjera, 24*, 1-43. http://marcoele.com/descargas/24/fernandez-evaluacion_aprendizaje.pdf

Ferreira, A. y Kotz, G. (2010). ELE-Tutor Inteligente: Un analizador computacional para el tratamiento de errores gramaticales en Español como Lengua Extranjera. *Revista signos, 43*(73), 211-236. https://dx.doi.org/10.4067/S0718-09342010000200002

García, M. (2024). ChatGPT: posibles aplicaciones y recomendaciones de uso en ELE. In *ELEUK ampliando horizontes: propuestas didácticas y avances en investigación* (pp. 121-139). Instituto Cervantes.

García Pujals, A. y Lasagabaster, D. (2019). El efecto de la evaluación y la retroalimentación en la autonomía, la motivación y el aprendizaje del español como L3. *Revista Española de Lingüística Aplicada, 32*(2), 455-485. https://doi.org/10.1075/resla.17050.gar

Hattie, J. y Timperley, H. (2007). The Power of *Feedback*. *Review of Educational Research*, *77*(1), 81-112.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T. y Kasneci, G. (2023). *ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences*, 103. https://doi.org/10.1016/j.lindif.2023.102274

López Mata, D. (2023). ChatGPT en la clase de preparación al DELE. Propuesta didáctica e impresiones de los estudiantes de ELE. *Revista Nebrija De Lingüística Aplicada a La Enseñanza De Lenguas, 17*(35). https://doi.org/10.26378/rnlael1735533

Mizumoto, A., Shintani, N., Sasaki, M. y Feng Teng, M. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics, 3*(2). https://doi.org/10.1016/j.rmal.2024.100116

Ranalli, J. (2021) L2 student engagement with automated *feedback* on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*. https://doi.org/10.1016/j.jslw.2021.100816

Ranalli, J., Link, S. y Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of *feedback* as part of argument-based validation. *Educational Psychology, 37*(1), 8-25. https://doi.org/10.1080/01443410.2015.1136407

Slamet, J. (2024). Potential of ChatGPT as a digital language learning assistant: EFL teachers' and students' perceptions. *Discoveries in Artificial Intelligence, 4*. https://doi.org/10.1007/s44163-024-00143-2

Xiao, Y. y Zhi, Y. (2023). An Exploratory Study of EFL Learners' Use of ChatGPT for Language Learning Tasks: Experience and Perceptions. *Languages, 8*(3). https://doi.org/10.3390