RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 62 (1), I Sem. 2024, pp. 13-38.

ISSN 0033-698X (impresa)
ISSN 0718-4883 (on-line)
https://doi.org/10.29393/RLA62-1IAMF30001

LA INTELIGENCIA ARTIFICIAL COMO HERRAMIENTA LEXICOGRÁFICA: ESTUDIO ANALÍTICO SOBRE EL RENDIMIENTO DE CHATGPT, COPILOT Y GEMINI EN UNIDADES LÉXICAS DEL ESPAÑOL[1]

ARTIFICIAL INTELLIGENCE AS A LEXICOGRAPHICAL TOOL: ANALYTICAL STUDY ON THE PERFORMANCE OF CHATGPT, COPILOT, AND GEMINI IN SPANISH LEXICAL UNITS

M.ª TERESA FUENTES MORÁN
Universidad de Salamanca, España
tfuentes@usal.es
https://orcid.org/0000-0002-1394-5535

FERMÍN DOMÍNGUEZ SANTANA
Universidad de La Laguna, España
fdominguez@ull.edu.es
https://orcid.org/0000-0001-5165-1282

CRÍSPULO TRAVIESO RODRÍGUEZ
Universidad de Salamanca, España
ctravieso@usal.es
https://orcid.org/0000-0002-0774-0728

RESUMEN

La popularización de las herramientas de inteligencia artificial generativa (IAG) ha creado escenarios inéditos de investigación en la metalexicografía, que se añaden a los ya derivados de las nuevas formas de acceso a la información léxica. En este contexto, el presente estudio analiza las entradas lexicográficas generadas por tres herramientas de inteligencia artificial generativa (ChatGPT, Copilot y Gemini) para un conjunto de veinticuatro términos, pertenecientes a seis categorías diferentes. Se plantea un enfoque

[1]Este artículo es parte del proyecto de I+D+i *PReLemma: Parámetros para recursos léxicos más accesibles (PID2022-137210OB-I00)*, financiado a través del Programa de Proyectos de Generación de Conocimiento del Ministerio de Ciencia e Innovación de España, MCIN/AEI/10.13039/501100011033.

descriptivo basado en tres aspectos: tipos de información, formas en que esta se presenta y coherencia con la información ofrecida por dos diccionarios monolingües de referencia (el *Diccionario de la lengua española* y el *Diccionario del español actual*). Se llevó a cabo un análisis basado en diez variables relativas a componentes formales del artículo lexicográfico y dos variables adicionales referidas a las fuentes de información y a la aparición de errores. Los resultados devuelven un comportamiento dispar para los elementos formales en cada una de las herramientas de IAG, con desempeños irregulares en todas las dimensiones tratadas. La valoración general de su rendimiento permite concluir que, actualmente, estos instrumentos no proporcionan una información estable, siendo aún esencial la labor humana para elaborar recursos léxicos. No obstante, su uso puede tener interés con un trabajo de posedición por parte de lexicógrafos o como base para aplicaciones didácticas orientadas a la mejora de la competencia léxica perseguida por los docentes, por ejemplo. Por último, cabe señalar que este estudio puede contribuir a la conformación de metodologías válidas para la incipiente investigación de estos nuevos entornos.

Palabras clave: Inteligencia artificial generativa, artículo lexicográfico, diccionarios, ChatGPT, Copilot, Gemini.

ABSTRACT

The popularization of generative artificial intelligence (GAI) tools has created unprecedented research scenarios in metalexicography, adding to those already derived from new ways of accessing lexical information. In this context, the present study analyses the lexicographical entries generated by three generative artificial intelligence tools (ChatGPT, Copilot, and Gemini) for twenty-four Spanish terms belonging to six different categories. A descriptive approach is proposed based on three aspects: types of information, ways in which it is presented, and similarity to the information provided by two monolingual dictionaries (*Diccionario de la lengua española* and *Diccionario del español actual*). An analysis was carried out based on ten variables related to the formal components of the lexicographical entry and two additional variables related to the sources of information and the presence of errors. Results show dissimilar behaviour for the formal elements in each of the GAI tools, with irregular performances across all dimensions addressed. The overall assessment of their performance allows us to conclude that, nowadays, these tools do not provide steady information, which makes human intervention still essential for developing lexical resources. Nevertheless, they might be valuable after post-editing by lexicographers or as a basis for didactical applications to improve lexical proficiency in language teaching, for instance. Finally, it should be noted that this study may help to shape valid methodologies for emerging research in these new environments.

Keywords: Generative artificial intelligences, lexicographical articles, dictionaries, ChatGPT, Copilot, Gemini.

Recibido: 02/05/2024 Aceptado: 10/06/2024

1. INTRODUCCIÓN

La necesidad de información léxica para las diversas actividades cotidianas o extraordinarias que desarrollamos no ha disminuido en los últimos tiempos. Aprendizaje y enseñanza, ocio y cultura, traducción y otras múltiples facetas de la comunicación intra e interlingüística exigen comprender, aprender, emplear viejas y nuevas palabras, y profundizar en su significado. Para ello, el diccionario podía proporcionarnos definiciones, ejemplos, indicaciones gramaticales o de uso, u otros datos relevantes. Y así, en efecto, acudíamos, entre otros recursos, a diccionarios, en los formatos y con las características que los avances científicos y tecnológicos nos han ido ofreciendo. La sociedad de la información puso a nuestro alcance gran cantidad de recursos que, aunque difusos en su mayor parte, parecían responder a nuestra exigencia de datos léxicos o por lo menos saciar nuestra curiosidad. Estos mismos avances tecnológicos llevan en gran parte de las ocasiones a que el usuario llegue a no tener conciencia de estar sirviéndose de una herramienta lexicográfica, ya que no *acude* propiamente a ella, sino que esta forma parte, podríamos decir de forma *subyacente*, de plataformas y servicios de alta tecnología, como asistentes de redacción, traductores automáticos, etc. Cambian no solamente los formatos, sino también, de manera más o menos consciente, las formas de *acceder* a los datos léxicos que se necesitan.

Paralelamente, la labor del lexicógrafo, como la de la mayoría de los oficios a lo largo de la historia, ha experimentado cambios muy significativos, por su evolución en relación con los avances tecnológicos, en especial gracias al crecimiento exponencial de las posibilidades de acceso a fuentes de información y a la automatización de los procesos (*cfr.* Rundell y Kilgarriff, 2011; Papadopoulou y Roche, 2019; Khan *et al.*, 2021; Rundell, 2024). Todavía algunos pueden recordar con nostalgia aquellas papeletas lexicográficas escritas cuidadosamente a mano, pero pocos las echarán realmente de menos.

En el marco de la denominada *cuarta revolución industrial* (Schwab, 2016), se reconocen transformaciones muy significativas, aunque para algunos esta etapa se caracterice más bien por la evolución acelerada de las innovaciones tecnologías desarrolladas en la etapa anterior. Recordemos que la denominada *tercera revolución industrial* suele situarse a partir de los años 60 y que se corresponde con la *revolución digital* y la *sociedad de la información globalizada* (Moll, 2021). En todo caso, uno de los hitos destacados de esta *cuarta revolución industrial* es la evolución y la (*cuasi*) generalización de la posibilidad de acceso a la inteligencia artificial generativa (IAG), lo que tiene repercusión en todos los ámbitos del saber, también en áreas vinculadas con las ciencias sociales o las humanidades, en ámbitos como la comunicación inter- e intralingüística o la adquisición de lenguas. En estas áreas –entendidas como transversales por su implicación en casi cualquier otra rama

del saber–, el imprescindible conocimiento léxico hace también imprescindible las *herramientas* que puedan proporcionarlo, sean estas más o menos semejantes a las conocidas o enteramente diferentes.

En este contexto, resulta legítimo plantear, en concreto, cuál es el rendimiento de las herramientas de IAG para proporcionar datos léxicos fiables, hasta qué punto es necesaria la intervención humana en esas tareas e incluso cuál es el papel que la lexicografía puede o debe desempeñar actualmente y en un futuro próximo, en pleno proceso de reconceptualización (Leroyer y Køhler Simonsen, 2020). Este trabajo propone un diseño para abordar este objeto de estudio, aplicándolo al ámbito de nuestro idioma. A partir de la revisión bibliográfica inicial, se plantea una metodología que comienza con el establecimiento de las variables de análisis, seguida de la selección de las unidades léxicas y los recursos evaluados. Tras la pertinente recogida y análisis de los datos, se ofrecen los resultados del estudio, vertebrados en función de la propia tipología de variables, de los que, finalmente, se extraen las conclusiones observadas.

2. INVESTIGACIÓN SOBRE INTELIGENCIA ARTIFICIAL GENERATIVA Y LEXICOGRAFÍA

Los trabajos más relevantes publicados en relación con el uso o la efectividad de las herramientas de IAG en el ámbito lexicográfico en los dos últimos años –aunque parciales, por propia naturaleza– aportan pautas metodológicas para abordar el estudio, así como unas primeras conclusiones descriptivas y analíticas de los resultados más allá de las discusiones y opiniones sobre si son o no necesarios los diccionarios o los lexicógrafos actualmente.

La herramienta utilizada es casi exclusivamente ChatGPT (versión GPT-3.5 del modelo) (Jakubiček y Rundell, 2023; Rees y Lew, 2024; Ortega-Martín *et al.*, 2023; Alonso Ramos, 2023; Arias-Arias *et al.*, 2024, entre otros). Para analizar su rendimiento, los autores seleccionan en primer lugar una serie de unidades léxicas que consideran representativas. La cantidad estudiada de estas unidades es muy variada –en consonancia, comprensiblemente, con la orientación del trabajo, las cuestiones que se plantean y, en consecuencia, las posibilidades de automatizar los procesos. Así, por ejemplo, Arias-Arias *et al.* (2024) trabajan a partir de 5 unidades en alemán y 5 en gallego, mientras que Ortega-Martín *et al.* (2023) utilizan 66.353 para probar la elaboración automatizada del *Spanish Built Factual Freectianary.*

También muy variados son los criterios de selección de estas unidades léxicas, criterios que en algunos casos no se explicitan. Por ejemplo, Lew (2023) selecciona 15 verbos de comunicación mientras que Jakubiček y Rundell (2023) califican las 99 unidades seleccionadas como "small set of very heterogeneous English

headwords" (p. 518).

Como es sabido, uno de los puntos críticos en este proceso es la decisión sobre cuáles son los *prompts*, es decir, la formulación de la instrucción o consulta más adecuada, que resultará más eficaz. De hecho, algunos autores optan por trabajar con más de una, de forma que se completen los resultados. En la tabla I se recoge una pequeña muestra de los *prompts* utilizados en cuatro trabajos.

Tabla I. Muestra de *prompts*.

Autor(es)	***Prompts***
Jakubiček y Rundell (2023)	¿Qué significa la palabra X? Generar una entrada de diccionario para X. Genere una entrada de diccionario para X que incluya posibles formas de la palabra, sentidos de la palabra, pronunciación, colocaciones, sinónimos, antónimos y ejemplos de uso.
Rees y Lew (2024)	Explica el sustantivo X (/verbo, etc.).
Ortega-Martín *et al.* (2023)	Genera en español una definición de la palabra X.
Rundell (2024)	Propone como posibles los siguientes: ¿Puedes definir la palabra W? Genera un artículo lexicográfico para W. Crea un artículo lexicográfico para W en que se muestren todos sus significados y usos en contextos diferentes.

Se observan tres tendencias generales: (1) las menos utilizadas son las consultas lingüísticas que podríamos valorar como naturales del tipo "¿Qué significa la palabra X?" (Jakubiček y Rundell, 2023), pero se priorizan los *prompts* con rasgos metalingüísticos, (2) ("explica el verbo x", "genera una definición para…") o bien (3) las consultas de carácter más lexicográfico (Jakubiček y Rundell, 2023).

Aquellos trabajos que abordan, más allá de la generación y análisis de respuestas, una comparación con diccionarios, lo hacen con fines prioritariamente estadísticos (Phoodai y Rikk, 2023, con *Oxford Advanced Learner's Dictionary* – OALD); desde la perspectiva del aprendizaje de lenguas (Rees y Lew, 2024, con *Macmillan English Dictionary*), o bien para disponer de una referencia para apoyar sus valoraciones (Jakubiček y Rundell, 2023, con *Oxford Dictionary of English – ODE–* y *Macmillan English Dictionary –MED*), Ortega-Martín *et al.,* 2023, con el *Diccionario de la lengua española (DLE)*, así como Lew, 2023, con *COBUILD*). Un paso más lo encontramos por ejemplo en el texto de Rees y Lew (2024), quienes

presentan un estudio experimental, con 43 participantes y utilizando 60 ítems, orientado a valorar la comprensión lectora en inglés.

En cuanto a los resultados obtenidos, destacamos que, por lo general, las definiciones se valoran como adecuadas, sobre todo cuando se trata de términos especializados, y se observa en estas una clara tendencia a mostrar patrones recurrentes (Rundell, 2024, pp. 8-9; Jakubíček y Rundell, 2023, pp. 526-528). Resulta representativo el estudio realizado por R. Lew (2023) quien, tras generar con IAG 15 entradas correspondientes a verbos de comunicación en inglés, las somete a evaluación a ciegas por 4 expertos junto con las mismas entradas extraídas de COBUILD. El resumen de los resultados de esta evaluación es el siguiente, aunque en la investigación se matiza con otras opiniones libres de los expertos que intervinieron en ella (figura 1):

Table 1 Central summary measures for expert ratings of definitions, examples, and entries created by AI and COBUILD lexicographers.

Element	Creator	Mean	Median	Mode
Definition	AI	3.6	Good	Good
Definition	COBUILD	3.9	Good	Good
Examples	AI	3.3	Passable	Passable
Examples	COBUILD	4.2	Good	Good
Entry	AI	3.2	Passable	Passable
Entry	COBUILD	3.8	Good	Good

Mean values are computed on a scale from 1 to 5.

Figura 1. Resumen de valoraciones de expertos sobre definiciones, ejemplos y entradas elaboradas con IA y por lexicógrafos de COBUILD (Lew, 2023, p. 4).

Se constatan carencias relevantes en la identificación de diferentes acepciones en unidades polisémicas, para muchos el mayor problema (Jakubíček y Rundell, 2023, pp. 525-526; Rundell, 2024, pp. 7-8). En este sentido, destaca el hecho de que se ignoren algunos sentidos, por lo general los figurados, y que una misma acepción se duplique, es decir, que se explique lo mismo en definiciones distintas separadas formalmente como diferentes acepciones; también en este terreno se registran algunas invenciones de significados. Rundell resume así estas deficiencias en los dos tipos de vaciados que presenta:

> The issues raised by these entries are symptomatic of problems found in most of the polysemous headwords in both samples: some meanings are duplicated; others are invented; and important meanings are omitted (in Sample B, *climate*

> has five 'senses' of the weather-related meaning, but none for the common metaphorical use, as in 'a climate of distrust'). On the basis of the two sample sets, it would be fair to conclude that ChatGPT performs best in handling single-sense words (especially technical terms), but is on shaky ground when confronted with even quite simple polysemous items of mainstream vocabulary. (2024, p. 8)

Por otro lado, son numerosos los errores en la categorización de clases de palabras (Jakubíček y Rundell, 2023, p. 527; Rundell, 2024). En cuanto a los ejemplos, se señala la tendencia a presentarse siguiendo patrones recurrentes (Rundell, 2024, pp. 8-9; Jakubíček y Rundell, 2023, pp. 526-528), pero especialmente gran cantidad de deficiencias que se manifiestan en ellos, lo que contrasta con aquellos que pueden extraerse con tecnologías actuales (Rundell, 2024, pp. 8-9).

Los casos en los que las acepciones ofrecen etiquetas (o marcas) para las acepciones, se presentan con frecuencia en forma de explicación, como en el caso que recogen Jakubíček y Rundell (2023, pp. 527-528) –y que califican como ejemplar, para la entrada correspondiente a *half-caste*, donde se indica que "It is now considered offensive and outdated and it is better to use terms such as 'mixed race' or 'multiracial' instead".

En el contexto de las aplicaciones de inteligencia artificial, se denomina *alucinación* al fenómeno en el que un modelo de lenguaje LLM (*Large Language Model*) "percibe patrones que son inexistentes o imperceptibles para los observadores humanos, creando resultados inesperados o incorrectos generados por los modelos de lenguaje" (Spinak, 2023). Es decir, por ejemplo, las invenciones que constaba Rundell (2024), como señalamos más arriba. Así pues, es previsible y comprobable que un porcentaje de los resultados obtenidos no puedan ser considerados correctos. Esto supone por supuesto un grave problema y no parece que vaya a ser solucionable a medio plazo. En efecto, algunos trabajos advierten alucinaciones o errores (Barrett, 2023) y constatan así mismo plagio y, en ocasiones, falta de actualidad en los datos.

3. DISEÑO DE UN ESTUDIO EXPLORATORIO APLICADO AL ESPAÑOL

3.1. Objetivos

En el contexto descrito, de marcado dinamismo en la aparición y transformación de las fuentes y herramientas tecnológicas, esta investigación no pretende profundizar en el dilema, que creemos superado, de posicionarse a favor o en contra de estos recursos, ni establecer una lista ordenada (y efímera) de los

mismos según el cumplimiento de determinadas características. Nuestro objetivo es realizar un estudio exploratorio y descriptivo sobre los resultados devueltos por estas herramientas cuando se les pide que faciliten información lexicográfica en español. Para el diseño metodológico partimos, por tanto, de una premisa de eventual complementariedad entre diccionarios e inteligencias artificiales generativas (IAG), no de sustitución o antagonismo entre recursos tan distintos en su esencia. Como fases de trabajo para lograr este objetivo, señalamos:

- Probar y adecuar –a partir de los trabajos reseñados más arriba– un modelo de análisis homogéneo y válido para distintas herramientas.
- Adaptar el modelo de interrogación a un contexto lexicográfico, abarcando distintos tipos de palabras.
- Extraer y clasificar la información obtenida.
- Analizar y valorar los puntos fuertes y debilidades en el rendimiento de las IAG.

3.2. Metodología

3.2.1. Configuración de prompts y variables de análisis

Siguiendo el modelo de trabajos anteriores (*cfr.* 2.), se diseñaron *prompts* que permitieran registrar el rendimiento de las IAG en el aporte de datos lexicográficos en lengua española. Para ello se utilizaron dos instrucciones:

a. Define la palabra *x*.
b. Crea una entrada de diccionario para la palabra *x*.

Se procuró evitar términos técnicos en la elaboración de estas instrucciones para no condicionar la respuesta de las IAG. Por otro lado, se optó por el uso de dos *prompts* y la recogida independiente de cada respuesta, para tener registro del diferente desempeño de las herramientas con una instrucción muy general (a) y una instrucción que condiciona el formato de la respuesta (b).

A la hora de determinar las variables de estudio, fue preciso elaborar un esquema propio que abarcara tanto los componentes formales –desde el punto de vista lexicográfico– en los que podíamos disgregar cada respuesta obtenida (los denominaremos *variables sobre componentes formales*), como aquellos que denominaremos *variables adicionales*, relevantes en el caso concreto que nos ocupa, y que hacen referencia aquí a las fuentes y al fenómeno de las alucinaciones. En las variables sobre componentes formales, se busca identificar y localizar la información correspondiente en un apartado concreto del artículo lexicográfico proporcionado por la herramienta, por ello las denominamos *formales*. Es decir, por ejemplo, la información sobre la morfología puede deducirse en ocasiones

de componentes como la definición o el ejemplo; en ese caso, no se trata de componentes formales, ya que no ocupan un lugar específico y explícito en el artículo lexicográfico. Con vistas a una mejor caracterización de los datos, ese fue el principal criterio diferenciador de las variables; seguidamente, para cada una de ellas se estimó una escala con los valores posibles.

Tabla II. Relación y clasificación de variables.

Variables sobre componentes formales	(1) Etimología (2) Sinonimia y antonimia (3) Información morfológica (4) Información combinatoria (5) Otras etiquetas (6) Definiciones (con referencia al número de acepciones y a la coincidencia con otros diccionarios) (7) Ejemplos de uso (8) Información complementaria (9) Información conversacional (10) Inclusión de imágenes
Variables adicionales	(A) Empleo de fuentes de información (B) Alucinaciones (o errores)

En la variable (4), *información combinatoria*, se computaron las unidades fraseológicas obtenidas, que, como es sabido, constituyen un componente determinante para la adecuada comprensión o producción textual del usuario de un diccionario (García Rodríguez, 2020; Penadés Martínez, 2017, 2024, entre otros). Por lo que respecta a (5), *otras etiquetas*, se recogió aquella información que el *Diccionario del Español Actual* adscribe a las siguientes categorías: "palabras anticuadas", "palabras de realidades lejanas en el tiempo o en el espacio", "vocabulario común activo o pasivo", "niveles de comunicación" y "vocabularios sectoriales. Nivel sociocultural. Nivel científico y técnico". A su vez, en (8), *información complementaria*, se midió la aparición de información no recogida en las otras variables. Por contraste, (9), *información conversacional*, recoge las apariciones más características, como veremos, del carácter comunicativo de las IAG y de giros conversacionales.

En el caso de la variable (A) *empleo de fuentes de información,* se registraron los casos en los que las IAG explicitaban las fuentes de las que obtuvieron datos, independientemente del origen de estos datos.

La mayoría de las variables respondían a valores dicotómicos, esto es, presencia o no de la información referida. En otros casos, podían tomar valores cuantitativos discretos (como fue el caso del número de acepciones o del número de informaciones combinatorias coherentes).

3.2.2. Selección de unidades léxicas

De manera simultánea, se abordó la elección de las palabras cuya definición serviría de base para el análisis. Una vez fijada la condición de contemplar varias categorías y en aras de evitar sesgos, se optó por un sistema aleatorio de selección, aprovechando para ello la posibilidad de consulta miscelánea aleatoria que ofrece el *Diccionario de la lengua española* en su versión gratuita en línea. Finalmente, las unidades léxicas incluidas son las que figuran relacionadas en la tabla III.

Tabla III. Listado de unidades léxicas utilizadas.

Categoría	**Unidades léxicas**
Sustantivos	*silla, exploración, pelo, aparcacoches*
Adjetivos	*púrpura, bizarro, inexpresivo, innato*
Verbos	*procrastinar, depreciar, intrincar, calcular*
Adverbios	*después, conjuntamente, quizás, casi*
Entradas recientes en DLE	*au pair, aquaplaning, baguette, balconing*
Léxico especializado	*angiosperma, fosforoscopio, litólogo, astenia*

3.2.3. Selección de las IAG

Como herramientas de las que obtener datos para comprobar el desempeño de las IAG, se seleccionaron tres instrumentos por su accesibilidad para el usuario común. En primer lugar, ChatGPT Plus permite el acceso a GPT-4, que es el LLM desarrollado por OpenAI. En segundo lugar, Copilot es un asistente de búsqueda propiedad de Microsoft que utiliza la tecnología de GPT-4 para procesar datos propios del modelo y actualizarlos a través de la búsqueda de fuentes en línea. En tercer lugar, Gemini es la herramienta de Google que permite el acceso al LLM propio desarrollado por la compañía.

3.2.4. Recogida y análisis de datos

El procedimiento seguido para extraer, registrar y analizar los datos siguió los siguientes pasos. En la obtención de datos, que tuvo lugar el 8 de marzo de 2024, se extrajeron, en primer lugar, las entradas de las unidades seleccionadas del *Diccionario de la lengua española* (*DLE*) (RAE, 2023, versión 23.7, edición en línea) y del *Diccionario del español actual* (*DEA*) (Seco *et al*., 2023, edición en línea). En segundo lugar, por lo que respecta a las IAG elegidas, se preguntó por todas las unidades léxicas seleccionadas con ambos *prompts*, procurando reiniciar la conversación tras cada interacción para evitar la contaminación por contexto de la respuesta previa. Como Copilot y Gemini permiten modificar la configuración de las respuestas obtenidas, se decidió, en el primer caso, obtener la respuesta más equilibrada de las tres opciones disponibles (*más creativo*, *más equilibrado* y *más preciso*) y, en el segundo, con configuración posterior, se registró la primera respuesta, sin atender a otras opciones (*más corta*, *más larga*, *más sencilla*, *más informal* o *más profesional*).

Para el procesamiento de las entradas de los diccionarios tradicionales y de las respuestas de las IAG se elaboró una base de datos *ad hoc* para organizar y clasificar la información en función de las distintas variables establecidas.

4. RESULTADOS

4.1. Variables sobre componentes formales

En relación con las variables del bloque de componentes formales, los datos obtenidos de las interacciones con las herramientas aparecen ilustrados en la figura 2, en las que se puede apreciar, de forma relativa, el rendimiento que cada IAG alcanzó sobre estos datos. Se contempló la aparición de los diversos tipos de información en las respuestas de las herramientas.

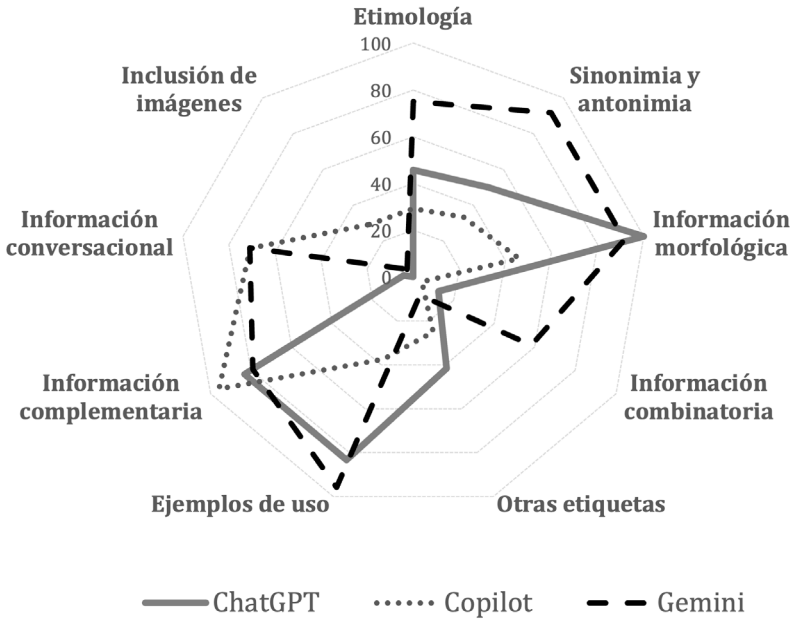

Figura 2. Rendimiento de las IAG en las variables sobre componentes formales.

Los resultados obtenidos en la consulta a los diccionarios y en la interacción con las IAG (segundo *prompt*) ponen de relieve la evidencia del carácter sincrónico, en el tratamiento de la etimología –variable (1)–, del *DEA* frente al *DLE* (18,18 % y 83,33 %, respectivamente). De hecho, como cabía esperar, dadas las características de este diccionario, las referencias al origen de las unidades consultadas en el *DEA* se refieren solo a la indicación de la lengua de procedencia en neologismos. Por su parte, la aparición de información etimológica en las respuestas de las IAG fue desigual (45,83 % en ChatGPT, 29,17 % en Copilot y 75 % en Gemini) y, en su gran mayoría, se correspondió con la información ya ofrecida por el *DLE*, aunque con una explicación más extensa sobre la formación de la palabra y el significado del étimo en el caso de la lengua latina. Así, por ejemplo, si en el diccionario se indica sobre *silla* "Del lat. *sella*" (RAE, 2023), ChatGPT completa "Del latín **sella**, diminutivo de **sedes**, que significa 'asiento'". Por el contrario, las IAG proporcionan información etimológica en unidades en las que el *DLE* no lo hace. Son los casos de *quizás* (en las tres herramientas), *inexpresivo* (ChatGPT y Gemini) y *aparcacoches* y *conjuntamente* (Gemini), aunque, en este último caso, solo se indica la formación de las palabras.

Por lo que respecta a la variable de aparición de sinónimos y antónimos –variable (2)–, ya que el *DEA* no ofrece este tipo de información, se comparó el rendimiento de las IAG con las unidades ofrecidas por el *DLE*, que las incluye desde su última actualización. Como resultado, observamos que frente al 75 % de presencia en el diccionario, las herramientas arrojaron un 50 % (ChatGPT), 33,33 % (Copilot) y 91,67 % (Gemini). Debido a que no fue el objetivo principal de este trabajo, no se estudió en profundidad la pertinencia como sinónimos o antónimos de las unidades ofrecidas por las IAG. Con todo, se detectaron fenómenos extraños. Por ejemplo, Copilot indicó *chófer* como sinónimo de *aparcacoches*, mientras que Gemini, para la misma unidad, propuso *gorrilla* y *párking*. Esta herramienta herró ofreciendo *calvicie* y *alopecia* como antónimos de *pelo* y, también, *pan de molde*, *panecillo* y *bollo* como antónimos de *baguette*. Además, incluyó como sinónimo de *después* la misma palabra en combinación (*después de, después que*). Este fenómeno también se observó en ChatGPT para *conjuntamente*. Por último, resulta relevante el comportamiento de esta herramienta con los sinónimos de *inexpresivo*, proponiendo como tales la traducción a varias lenguas (inglés, español, francés, italiano y alemán).

La presencia de información morfológica –variable (3)– fue absoluta en los casos del *DLE*, *DEA* y ChatGPT; casi completa en Gemini (91,67 %) y bastante menor en Copilot (45,83 %). La tónica general fue ofrecer información sobre la categoría de palabras; género y número para sustantivos y adjetivos, y carácter transitivo o intransitivo en el caso de los verbos. Estos datos, de forma general, se mantuvieron muy estables entre diccionarios e IAG, aunque se observó que, a diferencia de los diccionarios, donde se presentan esos datos parcialmente mediante abreviaturas, en las herramientas aparecieron sin abreviar. En el caso de *DLE* y el *DEA*, los verbos incluyen un modelo de conjugación y, en el primer caso, tablas con la conjugación completa. A este respecto, Copilot y Gemini ofrecieron esta información en los casos de *depreciar* e *intrincar* (las dos) y *procrastinar* y *calcular* (solo Gemini), mostrando los paradigmas de los tiempos de presente, pretérito imperfecto, pretérito perfecto simple, futuro (sin las tildes prescriptivas en el caso de *intrincar* con Gemini) y condicional de indicativo y, en ocasiones, presente de subjuntivo e imperativo. Otros fenómenos detectados fueron la falsa identificación de categoría de palabras en el caso de algunos adverbios. Por ejemplo, *después* fue identificado como preposición y conjunción por ChatGPT y Gemini. Por su parte, *conjuntamente* se categorizó como preposición por Copilot y Gemini y, además, en esta última herramienta, *quizás* también como preposición.

La variable *información combinatoria* –variable (4)– tuvo un resultado similar en los datos recogidos de los diccionarios (20,83 % en el *DLE* y 27,27 % en el *DEA*). Sin embargo, la inclusión de fraseologismos fue desigual en las respuestas ofrecidas por las IAG: inferiores en ChatGPT y Copilot (12,5 % y

4,17 %, respectivamente) y considerablemente superior en Gemini (65,95 %). Mientras que en los diccionarios estas unidades aparecen identificadas a través del formato (*DLE*), integradas entre las acepciones o marcadas como *loc* (*locución*) (*DEA*), en las IAG se registraron bajo diversas etiquetas en ChatGPT (*expresiones y usos figurativos* y *en combinaciones*, cada uno en una ocasión) y Gemini (*locuciones*, nueve veces; *refranes*, cuatro veces; *fraseología*, dos veces, y *expresiones*, una vez). Con todo, estas unidades ofrecidas por las herramientas no fueron coherentes en varias ocasiones, es decir, no incluyeron el lema del que se trataba el artículo. En el caso de *pelo* ChatGPT y Copilot proporcionaron tres ejemplos de los que uno no contenía la palabra. Por lo que respecta a Gemini, las incidencias fueron más frecuentes: *silla* (solo en tres fraseologismos incluyó la palabra de los treinta ofrecidos), *procrastinar* (cero de dos), *depreciar* (cuatro de ocho), *después* (tres de cinco), *quizás* (cero de tres) y *litólogo* (cero de dos).

En relación con la información identificada en el marco de la variable *otras etiquetas* –variable (5)–, se observó una presencia similar de estos datos en *DLE* (41,67 %), *DEA* (50 %) y ChatGPT (41,67 %), y menos en Copilot (25 %) y Gemini (8,33 %). Al igual que la información morfológica, fue frecuente la aparición mediante abreviatura en los diccionarios y sin abreviatura en las IAG. Por ejemplo, *Med* para una acepción de *púrpura* y *astenia* en los documentos lexicográficos frente a *En medicina* en ChatGPT y Copilot. Sin embargo, para este mismo (*astenia*), Gemini indicó *Med.*. Asimismo, ocurre con *angiosperma* (*Bot.* para *DLE* y *DEA*, pero *Botánica* para ChatGPT y Gemini) y *litólogo* (*Geol.* en *DLE* y *Geología* en Copilot). Un ejemplo mayor es *silla*, para la que se indica *hist* en una acepción del *DEA* y ChatGPT desarrolla "En contextos históricos o ceremoniales". No obstante, se observaron muestras desiguales como *balconing* (*Esp.* en *DLE*, *España* en Copilot, pero *Turismo* en ChatGPT). En ocasiones, además, ChatGPT ofreció estos datos en casos en los que los diccionarios no los incluían. Así fue en *exploración* ("En un contexto médico"), *bizarro* ("En desuso") y *después* ("uso coloquial").

En cuanto a las *definiciones* –variable (6)– se han sintetizado dos tipos de datos. Por un lado, se halló que la media del número de acepciones por artículo de los diccionarios tradicionales (2,71 acepciones/artículo para el *DLE* y 2,32, para el *DEA*) no difirió en gran medida de la que ofrecieron las IAG (2,33 en ChatGPT, 1,67 en Copilot y 2,08 en Gemini). Pese a no tratarse de lo más frecuente, el número de acepciones en algún caso fue muy dispar, como ocurrió con *pelo* (19 acepciones en el *DLE*, 9 en el *DEA*, 3 en ChatGPT, 10 en Copilot y 4 en Gemini). Por otro lado, como se puede observar en la figura 3, el número de acepciones presentadas por las herramientas de generación contrastaron considerablemente entre ChatGPT y Gemini, con respecto a Copilot. En este último caso, la totalidad de las acepciones ofrecidas por la IAG estaban recogidas en las obras

lexicográficas consultadas, mientras que en ChatGPT solo el 75 % y en Gemini el 76 %. Así, todas las acepciones proporcionadas por ChatGPT coincidían con las registradas en los diccionarios en doce palabras; en diez palabras una de las acepciones de la herramienta no concordaba y en dos entradas, dos de las voces. Por lo que respecta a Gemini, en diecisiete términos la totalidad de las acepciones entregadas se ajustaba con los diccionarios; en tres artículos una acepción no se consignó; en otros tres, dos acepciones y, por último, en un caso (*exploración*) tres acepciones no armonizaban.

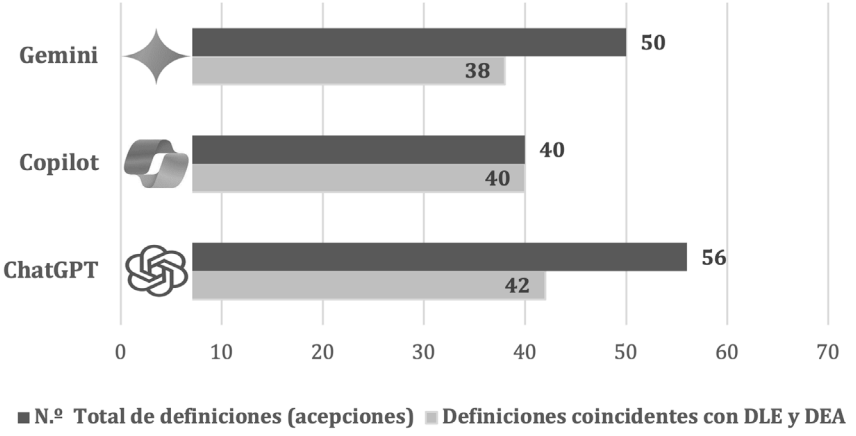

Figura 3. Rendimiento de las IAG en la variable 6 *Definiciones*.

Por otro lado, se advirtieron diversos fenómenos en el rendimiento de las herramientas con respecto a los diccionarios. Se llevó a cabo un análisis de la primera acepción de las fuentes y se detectó una alta incidencia de similitud entre la redacción ofrecida por los diccionarios y por las IAG. Así, ChatGPT reprodujo casi palabra por palabra la primera definición que proporciona el *DLE* para *exploración* y *conjuntamente* y el *DEA* para *silla*, mientras que hay similitud, pero no reproducción directa, de las primeras acepciones de *pelo, depreciar* y *balconing* (*DLE*). Con respecto a Copilot, sus primeras acepciones concordaron con el *DLE* en *exploración, aparcacoches, pelo, púrpura* (parcialmente), *bizarro, depreciar, intrincar, inexpresivo* y *conjuntamente*, aunque en estos dos últimos casos explicitó de forma directa la fuente bajo la fórmula "Definición según la Real Academia Española (RAE)" (Copilot). Por último, Gemini reproduce literalmente las

primeras acepciones del *DLE* en once voces (*silla, exploración, aparcacoches, pelo, depreciar, conjuntamente, quizás, casi* –aunque la distribuyó en varias acepciones–*, aquaplanning, baguette* y *astenia*) y de forma parcial en dos (*balconing* y *litólogo*).

El fenómeno más generalizado en la generación de entradas lexicográficas de las IAG fue la estrecha semejanza del contenido, con distinta redacción, en dos o más acepciones. Así ocurrió en ChatGPT para los términos *bizarro* (2 acepciones), *innato* (2), *calcular* (2), *después* (3), *conjuntamente* (3), *casi* (2), *aquaplaning* (2), *balconing* (2) y *astenia* (2). Con menor incidencia, se identificó este comportamiento en Copilot en los casos de *conjuntamente* (2 acepciones) y *astenia* (2). Por su parte, en Gemini sucedió con *casi* (3 acepciones) y *procrastinar* (2), para la que produjo una primera y segunda acepciones muy similares: "Diferir o aplazar una acción o tarea para un momento posterior" y "Retrasar voluntariamente la ejecución de algo, generalmente por desgana o pereza" (Gemini).

Otro fenómeno frecuente, solo detectado en ChatGPT y en Gemini, fue la adición de acepciones no recogidas en las obras lexicográficas ni en otras fuentes fiables consultadas. En este estudio no se juzga la pertinencia de estas acepciones, sino que solo se describe su aparición. Así, ChatGPT es la herramienta en la que, con mayor reiteración, se manifestó el comportamiento. Muestra de ello son la acepción tercera de *silla* ("En deportes ecuestres, se refiere al arte de montar a caballo, incluyendo las técnicas y habilidades necesarias para dirigir y controlar el animal"), la segunda de *exploración* ("En un contexto médico, se refiere al examen detallado del cuerpo de un paciente por parte de un profesional de la salud, con el fin de diagnosticar enfermedades o evaluar el estado de salud"), la segunda de *inexpresivo* ("Que carece de características destacadas, interesantes o estimulantes. Se aplica a objetos, lugares, obras de arte, etc., que no provocan una respuesta emocional o intelectual fuerte en el observador"), la tercera de *intrincar* ("Establecer una conexión o relación profunda y compleja entre personas o cosas, generando un vínculo que es difícil de deshacer"), la segunda de *quizás* ("Puede usarse para suavizar una afirmación o solicitud, haciéndola menos directa o categórica"), la segunda de *au pair* ("Programa o acuerdo cultural por el cual una persona se integra temporalmente en una familia en un país extranjero para ayudar con el cuidado de los niños y las tareas ligeras del hogar") y la tercera de *baguette* ("Moda: Bolsa o bolso de mano alargado y estrecho, similar en forma a la barra de pan francés, popular en diversas culturas de moda por su diseño único y su capacidad para complementar una amplia gama de estilos"). En su caso, Gemini crea nuevas acepciones para las palabras *silla* (acepción tercera: "En algunos juegos de naipes, cada uno de los grupos de cartas que se forman al repartirlas"), *exploración* ("Iniciación en la sexualidad", acepción tercera), *después* (acepción cuarta: "Consecuencia: Indica que algo es resultado de otra cosa" y acepción quinta: "Concesión: Indica que algo se concede a pesar de otra cosa") y

litólogo ("Perteneciente o relativo a la litología", acepción segunda).

Bajo el mismo criterio anterior (acepciones registradas que no aparecen en los diccionarios consultados), se identificó en dos voces la creación de acepciones por las IAG a partir de la acepción principal, pero con una especialización o precisión contextual mayor. Nuevamente, fue un comportamiento detectado en Gemini y ChatGPT. Para la palabra *exploración*, ChatGPT proporcionó como tercera acepción: "En sentido figurado, puede referirse a la investigación o indagación de aspectos inmateriales, como ideas, emociones o el potencial humano", mientras que Gemini indicó las siguientes acepciones segunda: "Reconocimiento minucioso de algo" y cuarta "Investigación científica, o su primer paso (el intento inicial de develar un conocimiento genérico sobre algún fenómeno)". Asimismo, para *depreciar* ChatGPT arrojó como segunda acepción "Menospreciar o desvalorizar a alguien o algo en cuanto a sus cualidades o méritos" y Gemini, también como segunda acepción, "Bajar la estimación o el aprecio que se tiene de alguien o algo".

Por último, un caso único, detectado en la respuesta de Gemini a *silla*, fue la acepción quinta: "Vehículo con asiento para una persona, a manera de caja de coche, y el cual, sostenido en dos varas largas, era llevado por hombres", que no aparece como definición en las obras lexicográficas consultadas, pero se identificó con la información del *DEA* para la forma compleja *silla de manos* "Vehículo formado por una caja con asiento para una pers. y con dos varas largas para ser portada en vilo por hombres" (Seco *et al.*, 2023).

Un último fenómeno único se observó en la respuesta de Copilot para *después*. Mientras que el *DLE* indica en las primeras tres acepciones diferentes matices del significado del término:

> 1. adv. Detrás o a continuación. […]
> 2. adv. Más tarde, o con posterioridad. […]
> 3. adv. Denota posterioridad en el orden, jerarquía o preferencia. (RAE, 2023)

Copilot, por su parte, amalgama estas tres acepciones en una sola: "Como adverbio, denota posterioridad temporal, espacial o jerárquica".

En las obras lexicográficas consultadas se registraron muy pocas palabras para las que se ofrecieran *ejemplos de uso* –variable (7)–. Así, el *DLE* solo lo indicó en tres ocasiones (*inexpresivo*, un ejemplo; *después*, siete ejemplos, y *casi*, dos). El *DEA*, por su parte, no ofrece datos a este respecto en ninguna ocasión, si no tenemos en cuenta los extractos del *corpus* que incluye la *Vista avanzada* en todos los casos (excepto para *fosforoscopio* y *litólogo,* que no están recogidas en este diccionario). Por el contrario, las IAG proporcionaron en numerosas ocasiones ejemplos de

uso. Gemini lo hizo para todos los términos buscados; ChatGPT, en veintidós ocasiones (excepto para *baguette* y *fosforoscopio*) y, en mucha menor medida, Copilot únicamente dio ejemplos de uso para *silla, pelo, aparcacoches, inexpresivo* (reprodujo el mismo ejemplo que el *DLE*)*, intrincar, calcular, después, conjuntamente, quizás* y *casi*. Aunque no se llevó a cabo un análisis exhaustivo de estos ejemplos en el presente estudio, podemos destacar algunos de los fenómenos observados.

a) En la totalidad de los casos, las oraciones generadas en los ejemplos de uso fueron coherentes, es decir, incluían el término buscado. No obstante, en ciertas ocasiones, aparecieron derivados. Así ocurrió, por ejemplo, en el caso de Gemini, que para *exploración* ofreció doce ejemplos de uso, entre los que incluyó: "Los exploradores españoles *exploraron* el Nuevo Mundo en el siglo XV" y "El gobierno anunció medidas para evitar la *depreciación* de la moneda" entre los diez ejemplos de uso que indicó para *depreciar* (destacados nuestros).
b) De forma general, los ejemplos proporcionados se relacionan, respectivamente, con cada acepción incluida en la definición bajo indicativos como *Ejemplos*, *Ejemplos de frases*, *Frase* o *Uso en oración*.
c) En una ocasión, los ejemplos estaban constituidos por citas (entre comillas) extraídas de documentos lexicográficos en línea, aunque sin que la IAG incluyera adecuadamente la referencia bibliográfica. Por ejemplo, Gemini ofreció para *bizarro* tres ejemplos que copia de la entrada de este adjetivo en el *Diccionario panhispánico de dudas*.
d) En dos casos, la herramienta proporcionó una *Frase célebre*, cuya autoría, de la que no se ofreció referencia concreta, no se pudo comprobar. Así, para *inexpresivo*, Gemini otorgó la cita "El rostro inexpresivo es el lienzo donde se pinta la hipocresía", que atribuyó a Jacinto Benavente; o la siguiente: "La procrastinación es el ladrón del tiempo", a Edward Young, en su respuesta para el verbo *procrastinar*.

Con respecto a la *información complementaria* –variable (8)–, es decir, explicaciones que proporcionaron las herramientas no atribuibles directamente a concreción del significado de los términos consultados, Gemini y ChatGPT incluyeron estos datos en el 79,17 % de las ocasiones y Copilot, en la totalidad de sus respuestas. Se incluyeron, en esta variable, las anotaciones sobre pronunciación (transcripciones fonológicas). Así, en el *DEA* se ofrece este tipo de información para los extranjerismos o falsos extranjerismos (*au pair, aquaplaning, baguette* y *balconing*); de la misma manera que Gemini, para *quizás* (y la forma *quizá*) en la respuesta al *prompt* 2 (P2).

Otro tipo de información de este tipo facilitada por las IAG son explicaciones

sobre causas, síntomas y tratamiento para afecciones como *púrpura* (Gemini, respuesta al *prompt* 1, P1), *procrastinar* y *astenia* (en estos dos últimos casos las tres herramientas en la respuesta a ambos *prompts*).

Otro grupo de datos los constituyeron origen o historia, forma y composición, usos y tipologías para *silla, pelo* y púrpura (con el significado de 'tinte') (solo ChatGPT, P1/P2), *baguette* (en Copilot y Gemini, P1/P2), y *angiosperma, fosforoscopio* y *litólogo* (en las tres herramientas y ambos *prompts*).

De la misma manera, se ofreció información sobre aspectos de uso lingüístico, como la diferencia de significado entre *bizarro* en lengua española e inglesa (ChatGPT, ambos *prompts*), la especialización del término *innato* dependiendo del área de conocimiento (ChatGPT y Gemini, ambos *prompts*), el uso de la forma *quizá* como alternativa a *quizás* (Copilot y Gemini, P2), la advertencia sobre la posible confusión entre *cuasi* y *casi*, la explicación sobre su uso en contextos coloquiales y formales (Gemini, P2) o la aclaración sobre el origen de la etimología de *calcular* (ChatGPT, P1). Por último, Gemini añadió una nota en su respuesta al P2 de *procrastinar* en la que afirmaba que "La forma *procastinar* también es válida, aunque la Real Academia Española recomienda la forma *procrastinar*".

De la misma manera, determinados términos resultaron más propensos a la aparición de este tipo de información. Así, la respuesta de Copilot al P2 de *calcular* dio lugar a una extensa explicación sobre la forma de calcular el porcentaje. Gemini, también en el P2, indicó que *quizás* aparece en la novela *El Quijote* de Cervantes y en la canción "Quizás, quizás, quizás", de Osvaldo Farrés. Las tres herramientas dieron datos sobre historia, condiciones de trabajo, características de los participantes y páginas web de referencia para *au pair* en ambos *prompts*. En las mismas condiciones se obtuvo información sobre causas, consecuencias y recomendaciones para evitar el *aquaplaning* e historia, características, consecuencias y campañas de prevención para el *balconing*.

Gemini tuvo la tendencia a ofrecer recursos adicionales, sugiriendo la consulta de términos relacionados con el buscado en tres ocasiones en el P2 y en una en el P1.

Por último, se observó como fenómeno generalizado en Copilot la inclusión de un párrafo final de síntesis en las respuestas a ambos *prompts* de doce términos (y en la respuesta de un *prompt* de ocho más), mientras que, en un número menor, Gemini tuvo el mismo comportamiento en la respuesta del P1 de tres voces, y ChatGPT en una única ocasión de la respuesta al mismo *prompt*.

En relación con la *información conversacional* –variable (9)–, incluida en las respuestas de las IAG, se manifestó una gran diferencia entre la presencia de este tipo de información en ChatGPT (solo en las respuestas de tres voces, el 12,5 %) frente a Copilot y Gemini (en ambos casos, en diecisiete ocasiones, el 70,83 %). La información de este tipo más frecuente fue derivada de la interacción

conversacional de las herramientas. Así, se registraron enunciados valorativos sobre la satisfacción de la respuesta. Copilot afirmó, para la respuesta al P2 de *silla*: "Espero que esta definición te sea útil. Si tienes alguna otra pregunta, no dudes en preguntar". Esta herramienta tuvo el mismo comportamiento cuatro veces (para el P2), una vez cada uno para ambos *prompts* y solo uno para el P1. Por su parte, Gemini añadió el enunciado en trece ocasiones (siempre para el P1). Por otro lado, se observaron enunciados introductorios para la respuesta ("A continuación, te presento una entrada de diccionario para esta palabra") y notas que advierten que el uso de una u otra definición depende del contexto en el que se use la palabra. Para el primer tipo, Copilot incluye el enunciado en tres ocasiones (dos para *pelo* y *púrpura*, P2; y una para *bizarro* ambos *prompts*), mientras que Gemini solo lo registró una vez (*bizarro*, P1). Para el segundo tipo, ChatGPT realizó la advertencia en dos términos (*púrpura* y *depreciar,* P1), Copilot también en dos voces (*aparcacoches* y *quizás*, P2) y Gemini una vez (*silla*, P1). En el comportamiento de Copilot se manifestó un patrón recurrente en la aparición de emoticonos que finalizaban las respuestas de trece términos (en tres casos, en la respuesta al P1; en seis, para el P2 y, en tres, para ambos *prompts*).

Por último, varios enunciados resultaron destinados a mostrar las limitaciones de las IAG o a ofrecer recomendaciones externas a la propia interacción. De esta manera, Copilot, en su respuesta al P2 de *calcular* recomendó usar calculadoras en línea para realizar cálculos específicos o consultar a un profesional de la salud para evaluar, tratar síntomas médicos y determinar causas adyacentes (en la respuesta a ambos *prompts* para *astenia*). Por su parte, Gemini ofreció enunciados de este tipo, siempre mayoritariamente en segunda persona:

- "Si quieres saber más, te recomiendo que consultes un diccionario especializado o busques información en internet" (*pelo*, P2).
- "Si tienes alguna pregunta sobre la púrpura como trastorno hemorrágico, te recomiendo que consultes a un médico" (*púrpura*, P1).
- "Si tienes dificultades para dejar de procrastinar, puede ser útil hablar con un terapeuta o consejero" (*procrastinar*, P1).
- "Si estás experimentando síntomas de astenia, es importante que consultes a un médico para que pueda determinar la causa y ofrecerte el tratamiento adecuado" (*astenia*, P1).
- "Para obtener información más específica, se recomienda consultar con un especialista en litología" (*litólogo*, P2).

En cuanto a inclusión de imágenes –variable (10)–, Copilot incluyó imágenes en siete ocasiones y Gemini solo en una. Copilot erró en la selección, por ejemplo,

cuando la imagen que ofreció para ilustrar *después* solo mostraba la palabra *adverb* del inglés.

Figura 4. Imagen ofrecida por Copilot para *después*.

4.2. Variables adicionales

La tabla IV muestra el ámbito de aparición de esta tipología de datos en las respuestas de las IAG a ambos *prompts*.

Tabla IV. Rendimiento de las IAG en las variables adicionales.

	ChatGPT	Copilot	Gemini
Empleo de fuentes de información	0 %	100 %	75 %
Alucinaciones o errores	58,3 %	45,8 %	79,2 %

El uso de *fuentes* de información –variable (A)– es intrínseca a la propia naturaleza de las IAG. Aun así, lo que evidenciaron la coincidencia de la redacción de las primeras acepciones o la copia de algunos ejemplos de uso es la manifiesta cercanía que en ocasiones se mantiene con algunas fuentes de información. En el caso de las obras lexicográficas, tanto el *DLE* como el *DEA* utilizan, como es sabido, corpus textuales de elaboración propia. Sin embargo, no todas las herramientas de IAG explicitaron las fuentes utilizadas. De hecho, el comportamiento es muy desigual entre las tres estudiadas. ChatGPT en ningún caso indicó sus fuentes; Gemini, en el 75 % de las consultas, y Copilot, en la totalidad. De las fuentes explicitadas, el recurso lexicográfico más utilizado fue la versión en línea del *DLE* (veinte ocasiones en Copilot y trece en Gemini), seguido del *Cambridge dictionary* (cinco veces en Copilot), el *Diccionario panhispánico de dudas* (tres consultas en cada caso), el *Collins dictionary* (*silla* en Copilot), el *Diccionario médico de la Universidad de Navarra* (*astenia* en ambas herramientas) y el *Diccionario del español de México de El Colegio de México* (*bizarro* en Copilot). Como otras fuentes fiables de información lingüística, también se registraron el portal FundéuRAE

(utilizado en dos ocasiones por las dos IAG) y el portal de consultas lingüísticas de la RAE (para *quizás* en ambos casos).

Con todo, también se registró la incidencia de otras fuentes explicitadas. Por ejemplo, para *pelo, procrastinar, au pair* y *aquaplaning* (Gemini) y para *balconing* y *angiosperma* (Copilot). Estas otras fuentes de información fueron, por orden de aparición, diarios y portales de información generalista, Wikipedia, Wordreference y otras páginas de diccionarios en línea (Thefreedictionary.com, SpanishDictionary.com, Significado.com, Significados.com, Definicion.de, Definicion.com).

Copilot, en ciertas ocasiones, explicitó otro tipo de fuentes. Por ejemplo, la página web del Ayuntamiento de Silla (Valencia) para la palabra *silla*; para *aparcacoches* los portales Minijuegos.com y Disneyplus.com; la página web de la Cooperativa Agrícola San Isidro (Almería) para *casi* o páginas de información sobre las partes de un diccionario y de una entrada lexicográfica[2].

En cuanto a la variable (B), alucinaciones o errores, en los párrafos anteriores se han ido indicando algunos de ellos en las respuestas obtenidas de las IAG que guardan relación con el tipo de información incluido en cada variable. Sin embargo, se observaron algunos otros comportamientos que deben reseñarse.

Para el caso de ChatGPT, además de los aspectos mencionados con anterioridad, se observó una falta de concordancia en un ejemplo de uso proporcionado en el P2 de *pelo*: "El pelo de este [*sic*] brocha es ideal para pintura al óleo". Además, en esta misma respuesta, el ejemplo de uso para la primera acepción puede considerarse excesivamente extenso:

> El pelo de un gato no solo le proporciona aislamiento térmico sino también sensibilidad ante la presencia de estímulos cercanos gracias a los folículos pilosos dotados de terminaciones nerviosas. (ChatPGT)

Este mismo comportamiento (ejemplo de uso de gran extensión) se experimentó en Gemini en la respuesta al P2 de *au pair*. Esta herramienta, a su vez, respondió en lengua inglesa en dos ocasiones (*calcular* y *después*), incluyó caracteres japoneses en un ejemplo de uso para *litólogo* ("La litóloga española, María Dolores**なのですが,** ha realizado importantes investigaciones sobre las rocas volcánicas"), definió *intrínseco* cuando se le pidió (P1) la definición de *intrincar* y ofreció enlaces rotos a información adicional en línea en la palabra *inexpresivo*. Para *quizás*, indicó que se conjuga (*sic*) como un adverbio y en un apartado que denomina *Usos*, proporcionó información contradictoria:

[2]Entre las páginas consultadas están las siguientes: https://ccfprosario.com.ar/como-hacer-una-entrada-de-diccionario/ y https://www.unprofesor.com/lengua-espanola/el-diccionario-y-sus-partes-264.html.

- "Se utiliza para expresar duda, probabilidad o incertidumbre".
- "También se puede utilizar para dar por seguro o por cierto algo".

Asimismo, para *conjuntamente*, proporcionó información que puede caracterizarse como superflua, como que "Se puede utilizar tanto en oraciones afirmativas como negativas", o errónea, como que se trata de una locución adverbial o que palabras derivadas de este término son "conjunción, conjunta, conjunto".

5. CONCLUSIONES

Una vez revisados los aspectos que se han presentado, no podemos mantener el optimismo que algunos promulgaban hace meses (de Schryver, 2023). Las herramientas de IAG no nos han proporcionado información fiable en los contextos establecidos en este trabajo. Bien es cierto que esta tecnología puede mejorar gracias, entre otros avances, al aumento del volumen de datos con los que se entrena (Rundell, 2024, p. 9). Por otro lado, hay margen para ajustar las instrucciones y, quizá, mejorar los resultados, pero, a corto plazo, parece que sería más eficaz para el lexicógrafo o para el docente, que para un usuario general –es decir, aquel cuya consulta está motivada por una laguna de conocimiento léxico que desea remediar, destinatario genuino de los compendios lexicográficos.

No es de obviar que las investigaciones precedentes y la desarrollada en este trabajo, cada una con su particularidad, constituyen algunos acercamientos que pueden ir conformando y validando una nueva metodología complementaria para abordar este nuevo escenario.

A día de hoy, cabe reconocer que si bien estas herramientas pueden tener su utilidad con fines lexicográficos, la incertidumbre de las respuestas es mucho mayor que la de herramientas ya existentes. Parece que, por lo menos a medio plazo, el papel humano en la elaboración de recursos léxicos fiables es indispensable. Se trata de la tarea denominada *posedición*, que ya se venía recomendando y desarrollando antes de la popularización de la IA (p. ej. Steurs *et al.*, 2020).

Este estudio tenía como propósito conocer algo más profundamente la IAG como herramienta lexicográfica. Las limitaciones están claras. A las que son características de la extensión de un trabajo de esta índole, como las restricciones de los *prompts* empleados, de la extensión del corpus y de los parámetros que pueden estudiarse en profundidad, se le suma especialmente la inestabilidad de las respuestas obtenidas –que pueden variar más o menos significativamente si se repiten en otro momento. Sin embargo, la inclusión de tres herramientas y la amplitud de las variables estudiadas pueden aportar nuevo conocimiento del objeto de estudio para el español. Se abren nuevas vías de investigación de índole

teórico y de carácter aplicado, como el planteamiento de pautas de posedición para el uso de las herramientas con fines lexicográficos –sean estos para uso humano directo o en el marco subordinado a aplicaciones–, o de planteamientos didácticos que profundicen en el análisis crítico para la mejora de la competencia léxica del aprendiz.

REFERENCIAS BIBLIOGRÁFICAS

Alonso Ramos, M. (2023). El papel de ChatGPT como lexicógrafo. En C. Garriga Escribano, S. Iglesia Martín, J. A. Moreno y A. Nomdedeu Rull (Eds.), *Lligams. Textos dedicats a Maria Bargalló Escrivà* (pp. 15-26). Universidad Rovira i Virgili.

Arias-Arias, I., Domínguez, M. J. y Valcarcel, C. (2024). Der Effizienz- und Intelligenzbegriff in der Lexikographie und künstlichen Intelligenz: kann ChatGPT die lexikographische Textsorte nachbilden? *Lexikos*, *34*, 51-76. https://doi.org/10.5788/34-1-1879

Barrett, G. (2023, mayo 31-junio 3). *Defin-O-Bots: Challenging A.I. to Create Usable Dictionary Content* [Conference presentation]. 24th Biennial Conference of the Dictionary Society of North America, Boulder, CO, USA.

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, *36*(4), 355-387. https://doi.org/10.1093/ijl/ecad021

García Rodríguez, J. (2020). Hacia un diccionario electrónico de fraseología bilingüe en español y catalán: reflexión en torno a sus funciones y las percepciones de los usuarios. *RLA. Revista de Lingüística Teórica y Aplicada, 58*(1), 13-36. https://doi.org/10.29393/RLA58-1JGHD10001

Jakubíček, M. y Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? En M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček y S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference* (pp. 518-533). Lexical Computing CZ s.r.o. Disponible en https://elex.link/elex2023/proceedings-download/

Khan, M. Y., Qayoom, A., Nizami, M. S., Siddiqui, M. S., Wasi, S. y Razzi, S. M. K. (2021). Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. *Complexity*, art. 2553199. https://doi.org/10.1155/2021/2553199

Leroyer, P. y Køhler Simonsen, H. (2020). Reconceptualizing Lexicography: The Broad Understanding. En Z. Gavriilidou, M. Mitsiaki y A. Fliatouras, (Eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion.*

Volume 1 (pp. 183-192). Ljubljana University Press. Disponible en https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p183-192.pdf

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, *10*, art. 704. https://doi.org/10.1057/s41599-023-02119-6

Moll, I. (2021). The Myth of the Fourth Industrial Revolution. *Theoria, 68*(2), 1-38. https://doi.org/10.3167/th.2021.6816701

Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Armenteros, J. C., Álvarez, J. y Alonso, A. (2023). *Spanish Built Factual Freectianary (Spanish-BFF): the first AI-generated free dictionary*. arXiv. https://doi.org/10.48550/arXiv.2302.12746

Papadopoulou, M. y Roche, Ch. (2019). Twinning Classics and A.I.: Building the new generation of ontology-based lexicographical tools and resources for Humanists on the Semantic Web. *TwinTalks@DHN*. Disponible en https://api.semanticscholar.org/CorpusID:162183429

Penadés Martínez, I. (2017). Arbitrariedad y motivación en las colocaciones. *RLA. Revista de Lingüística Teórica y Aplicada, 55*(2), 121-142. http://dx.doi.org/10.4067/S0718-48832017000200121

Penadés Martínez, I. (2024). Los diccionarios fraseológicos. En S. Torner, P. Battaner e I. Renau (Eds.), *Lexicografía hispánica. The Routledge Handbook of Spanish Lexicography* (pp. 537-550). Routledge/Taylor & Francis Group.

Phoodai, Ch. y Rikk, R. (2023). Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework. En M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček y S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference* (pp. 345-375). Lexical Computing CZ s.r.o. Disponible en https://elex.link/elex2023/proceedings-download/

Real Academia Española (2023). *Diccionario de la lengua española* (23.ª ed.). [Versión 23.7 en línea]. Disponible en https://dle.rae.es

Rees, G. P. y Lew. R. (2024). The Effectiveness of OpenAI GPT-Generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically Orientated Reading Task. *International Journal of Lexicography, 37*(1), 50-74. https://doi.org/10.1093/ijl/ecad030

Rundell, M. (2024). Automating the creation of dictionaries: are we nearly there? *Humanising Language Teaching*, *26*(1). Disponible en https://www.hltmag.co.uk/feb24/automating-the-creation-of-dictionaries

Rundell, M. y Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? En F. Meunier, S. de Cock, G. Gilquin y M. Paquot (Eds.), *A Taste for Corpora. A tribute to Professor Sylviane Granger* (pp. 257-281). John

Benjamins. https://doi.org/10.1075/scl.45.15run

Schwab, K. (2016). *Fourth industrial revolution. Founder and executive chairman, World Economic Forum*. Crown Business Publishing.

Seco, M., Andrés, O. y Ramos, G. (2023). *Diccionario del español actual*. Disponible en https://www.fbbva.es/diccionario

Spinak, E. (2023, diciembre 20). ¿Es que la Inteligencia Artificial tiene alucinaciones? *SciELO en Perspectiva*. Disponible en https://blog.scielo.org/es/2023/12/20/es-que-la-inteligencia-artificial-tiene-alucinaciones/

Steurs, F., Schoonheim, T., Heylen, K. y Vandeghinste, V. (2020). *The Future of Academic Lexicography -- A White Paper* [White paper]. Versión 1.2. Instituut voor de Nederlandse taal. Disponible en www.ivdnt.org