ChatGPT en el país de los lingüistas

Los grandes modelos del lenguaje han sorprendido a muchos al ser capaces de redactar largos textos coherentes, pero una mirada atenta guiada por la Lingüística permite afirmar que no son modelos de la competencia lingüística humana, sino, por ahora, “loros estocásticos”.

La llegada de los grandes modelos del lenguaje como ChatGPT (LLM, por *Large Language Models*, es la sigla inglesa que se impuesto) ha supuesto un cañonazo en varios ámbitos sociales: la tecnología, la ética y, naturalmente, la lingüística. ¿Hemos cambiado de paradigma, quizá de era?

El debate surgió inmediatamente y ha habido reacciones para todos los gustos. Citemos algunas de las más paradigmáticas: Bender *et al.* muestran una gran cautela, llaman la atención sobre el coste energético de alimentar los LLM y sobre los posibles sesgos ideológicos que pueden generar (dos aspectos muchas veces ausentes en una discusión más basada en el “milagro” de la máquina que habla) y muy claramente afirman que estas IIAA no son modelos de la competencia lingüística de los seres humanos. Chomsky *et al.*, en un artículo de opinión en el New York Times, son apocalípticos y llegan a invocar el concepto de la “banalidad del mal”, acuñado por Hannah Arendt como subtítulo de su libro *Eichmann en Jerusalem*. Esta acusación tan seria –de verdad la artillería pesada: se habla del holocausto- se hace a raíz de la falta de ética de la inteligencia artificial (IA), no de su capacidad lingüística, pero muestra que la IA va a estar en el centro del debate intelectual durante los próximos años. En el lado de los entusiastas, destaca Piantadosi con un trabajo que ha producido un intenso debate. En él se afirma que los LLM vienen a constituirse en la derrota de las tesis innatistas chomskianas sobre el lenguaje humano, que han sido esenciales en el debate desde los años 50, y que los LLM sí pueden considerarse modelos de la competencia de los hablantes, contrariamente a lo que sostienen Bender *et al.*, Chomsky *et al.* y muchos otros. Se puede comprobar, pues, que las opiniones sobre estas nuevas IIAA van de un extremo a otro en muchos casos de manera más programática que empírica.

En cualquier caso, hay una idea de Piantadosi que me parece esencial: ChatGPT y los otros LLM han puesto el lenguaje en el centro del debate sobre la IA y en el centro de la actividad tecnológica que va a condicionar nuestro futuro como seres humanos. Ello no es más que el eco de la idea chomskiana de que el lenguaje define a la especie humana: el *homo loquens*.

No se conoce ningún sistema de comunicación animal parecido al lenguaje humano

Si el lenguaje va a estar en el centro del debate, no está de más, pienso, consultar la opinión de los lingüistas sobre la cuestión y pedirles no grandes declaraciones epistemológicas, sino algún dato empírico sobre lo que de verdad hacen los LLM. En 70 años de gramática generativa deberíamos haber aprendido lo suficiente para formular preguntas incisivas a una IA lingüística.

En estas líneas quisiera contribuir a la reflexión exponiendo algunos fenómenos bien conocidos de los gramáticos y de esta manera explicar por qué los LLM han merecido el apodo de “loro estocástico”, repetidor probabilístico, un término que aparece en el trabajo citado de Bender *et al.* y que está destinado a quedarse, pienso.

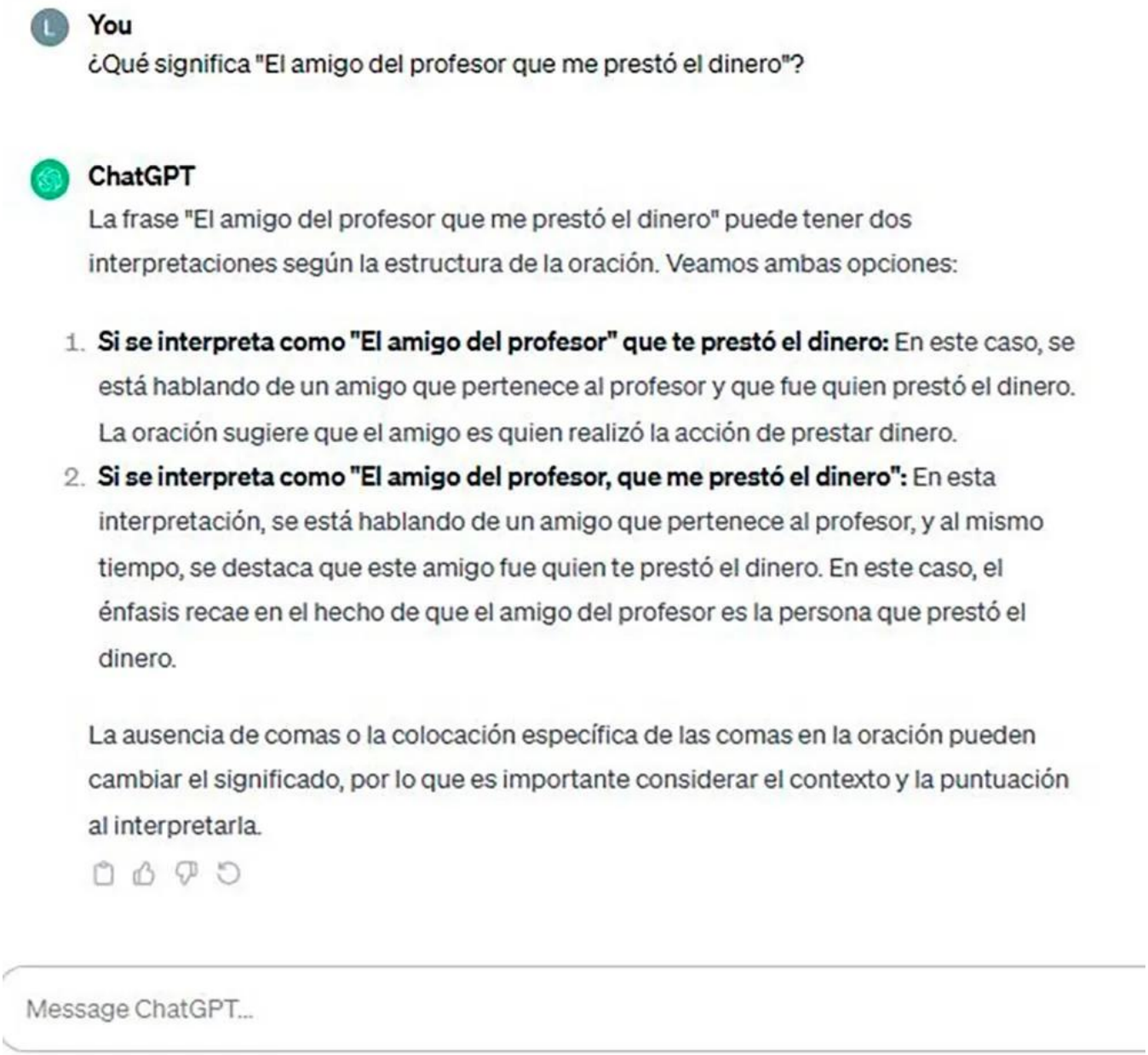
En primer lugar, vamos a proponerle un ejemplo básico a ChatGPT: *libros y muebles viejos*. La respuesta de ChatGPT a la pregunta de qué significa la secuencia es la siguiente:

Fuente: Luis García Fernández

El ejemplo es conocido porque constituye un caso básico de ambigüedad estructural con solo cuatro palabras. Las ambigüedades estructurales se diferencian de las léxicas en que no es un elemento concreto del léxico el que produce la ambigüedad como en *Ha desaparecido el gato*, donde puede haber desaparecido el gato

animal o el gato hidráulico. Las ambigüedades estructurales nacen de la posibilidad de asignar dos análisis sintácticos, dos segmentaciones, a una secuencia. Puede observarse que la interpretación a la que accede ChatGPT es aquella en que el adjetivo *viejos* modifica a los dos nombres coordinados *libros* y *muebles*, de modo que de ambos se predica la propiedad de ser viejos. Pero no puede acceder a la otra representación, en la que viejos solo modifica a muebles y, por lo tanto, no se puede concluir nada sobre la antigüedad de los libros.

Propongamos a la IA un segundo ejemplo llamado *El amigo del profesor de matemáticas que me prestó el dinero*:

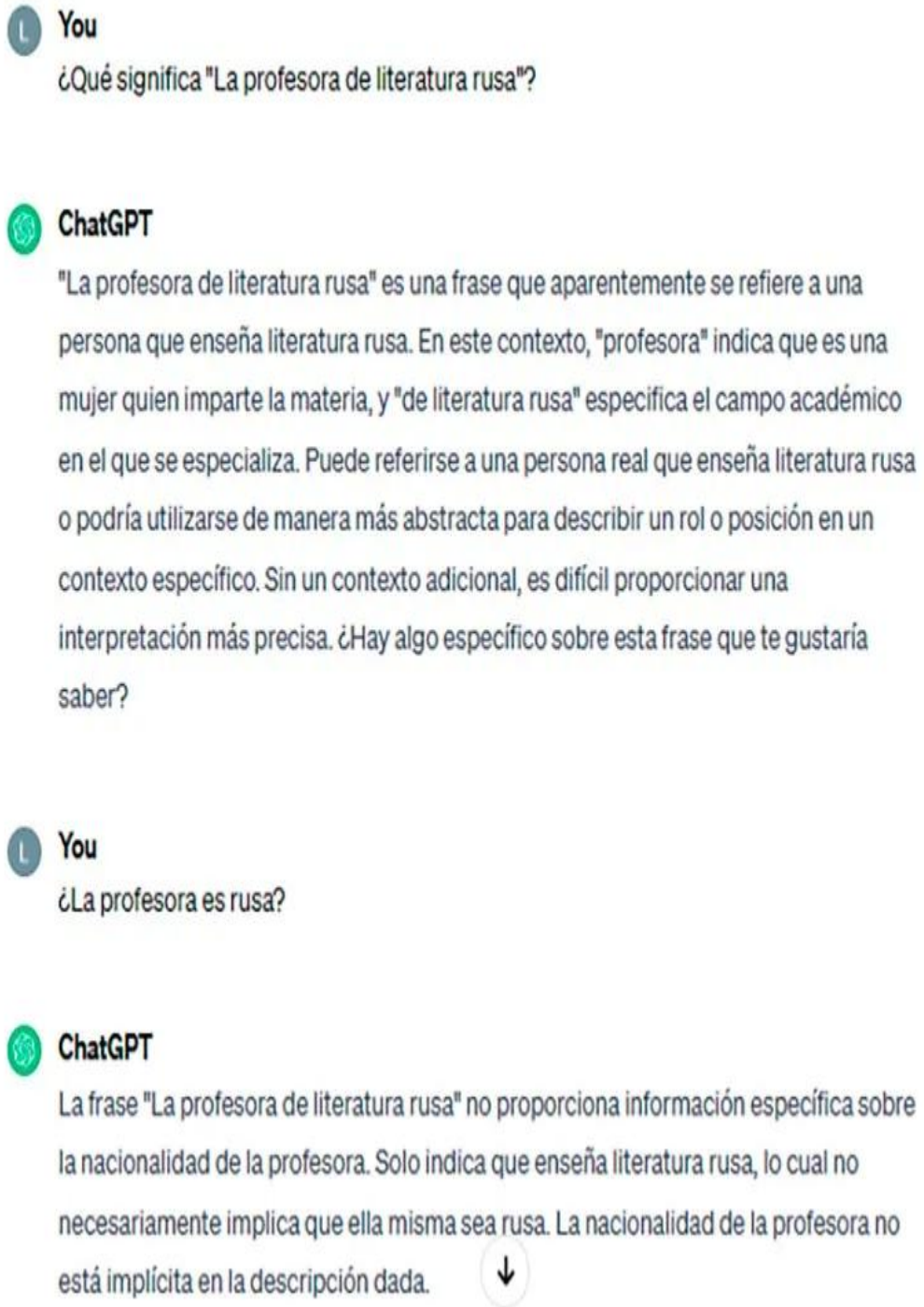


Fuente: Luis García Fernández

Simplificando las cuestiones técnicas, el segmento *que me prestó el dinero* puede tener dos antecedentes, es decir, dos elementos de los que tomar la referencia, el amigo o el profesor de matemáticas; por esta razón, puede que me prestase dinero el amigo o el profesor de matemáticas. Obsérvese que ChatGPT accede a la interpretación en la que el amigo me prestó el dinero, pero no a aquella en la que me presta el dinero el profesor. No es este, evidentemente, el único problema, aunque nos centremos en él. De ningún modo, el amigo pertenece al profesor, pero dejamos de lado esa cuestión.

El último ejemplo proviene de una conferencia de José Luis Mendívil-Giró en la Universidad Complutense: *la profesora de literatura rusa*. Esta secuencia es otra vez ambigua entre una interpretación en la que la literatura es rusa y otra en que la profesora es rusa. De nuevo, ChatGPT solo tiene acceso a la primera, pero

no a la segunda. Es decir, esta IA es incapaz de analizar la secuencia de dos maneras distintas y, por lo tanto, de manejar las dos interpretaciones: tanto aquella en la que la profesora, de cualquier nacionalidad, imparte literatura rusa, como la otra, donde la profesora de literatura de una lengua que no se nombra es rusa:



Fuente: Luis García Fernández

De estos ejemplos simples, se pueden extraer conclusiones interesantes. La primera es que ChatGPT no procesa las secuencias como una mente humana, a pesar de que esta IA puede redactar textos largos sorprendentemente bien escritos y coherentes. Piantadosi muestra como ejemplo un relato de ChatGPT elaborado a partir de la consigna: “Escribe una breve historia explicando cómo una hormiga podría hundir un portaaviones”. El resultado es verdaderamente sorprendente e invitamos al lector a hacer una prueba análoga.

¿Cómo es posible que ChatGPT componga un texto largo y coherente, pero no sea capaz de identificar las sencillas ambigüedades de los ejemplos que hemos propuesto? Vamos a intentar explicarlo desde el punto de vista de la lingüística teórica.

Chomsky (1986) establece la diferencia entre Lengua-E(xteriorizada) y Lengua-I(nteriorizada). Para Chomsky, la lengua-E se corresponde con la visión estructuralista de Leonard Bloomfield, según la cual, una lengua es la totalidad de las proferencias que se pueden hacer dentro de una comunidad lingüística. En este marco, una gramática es una colección de enunciados descriptivos compuestos de una representación fónica y una semántica: todos los textos españoles que se pueden encontrar en una biblioteca o en una base de datos, por ejemplo.

A este concepto de Lengua-E, Chomsky le opone el de Lengua-I. Para este autor, que una persona hable una lengua no parece implicar que conozca un conjunto infinito de oraciones, o de pares sonido-significado considerados en cuanto a su extensión, o un conjunto de actos o conductas; más bien implica que la persona sabe lo que hace que el sonido y el significado se relacionen de una forma específica en esa determinada lengua, que es, como dice Chomsky, el significado que comúnmente se atribuye a *Juan sabe francés*. La Lengua-I es una propiedad de la mente-cerebro de los seres humanos, no un conjunto de datos externalizados y recogidos en bases de datos por inmensas que estas sean. Esta propiedad humana, la Lengua-I, hace que los seres humanos sean distintos del resto de los seres vivos. No se conoce ningún sistema de comunicación animal parecido al lenguaje humano.

ChatGPT concatena secuencias de signos, pero no tiene reglas gramaticales

La cuestión es que la lengua que diseña ChatGPT es externalizada en el sentido de que puede producir secuencias idénticas a las que produce un hablante de una lengua natural, pero no es interiorizada en el sentido de que carece de las reglas gramaticales que maneja ese hablante. La producción de secuencias lingüísticas carece de reglas gramaticales; nace de un cálculo probabilístico sobre el análisis de millones de textos, de tal manera que el modo en que ChatGPT produce o “entiende” *libros y muebles viejos*, *el amigo del profesor de matemáticas que me prestó el dinero* o *la profesora de literatura rusa* es completamente diferente al modo en que las produce y entiende (ahora sin comillas) un hablante de español. ChatGPT concatena secuencias de signos, pero no tiene reglas gramaticales.

En teoría de conjuntos se sabe que cualquier conjunto se puede definir intensional o extensionalmente. Si tomamos el conjunto de los Reyes Magos lo podemos definir intensionalmente como los sabios que vinieron de Oriente a adorar al Niño Jesús o extensionalmente nombrándolos: Gaspar, Melchor y Baltasar. Podría pensarse que vienen a ser lo mismo, pero son radicalmente diferentes. Piénsese en el conjunto de los

españoles: una definición intensional cabe en una frase, pero una extensional sería el censo completo e incluiría millones de nombres. Una gramática de una lengua natural no solo es interiorizada en el sentido de Chosmky (1986), sino que es intensional, porque proporciona reglas de formación y no contiene secuencias concretas de palabras. Lo que produce ChatGPT es puramente extensional y lo hace con la misma capacidad con la que una máquina podría dar el censo completo de los españoles, pero ChatGPT no contiene ni una sola regla gramatical, de modo que puede elaborar historias complejas calculando la unidad más probable que sigue a cada una de las anteriores, pero tropieza irremediablemente con la ambigüedad estructural de la secuencia *libros y muebles viejos*. En este sentido, tienen razón Bender *et al.*: ChatGPT es un loro estocástico, es decir, un loro con una memoria prodigiosa, pero que no es capaz de analizar como un ser humano secuencias muy simples.

En realidad, el conjunto de secuencias bien formadas que define una gramática interiorizada o intensional y el conjunto de secuencias generadas por ChatGPT son coincidentes solo en parte. Para una gramática interiorizada, el modo en que se obtienen las dos interpretaciones que se asignan a *libros y muebles viejos* es distinto; cada interpretación tiene una estructura diferente. ChatGPT solo accede a una de las interpretaciones y lo hace de un modo diferente a cómo lo hace una gramática interiorizada. De alguna manera, ChatGPT es un gigantesco espejismo: es algo que, *prima facie*, parece lenguaje humano, pero que no lo es estrictamente hablando.

Nada de lo que hemos dicho quita mérito al diseño de los LLM, que son y serán extraordinariamente útiles, sino que explica que no son modelos de las lenguas naturales que conocen los hablantes, sino modelos aproximados de las secuencias que externalizan. No más, pero tampoco menos. Dice Piantadosi, y muy probablemente tiene razón, que la revolución de los LLM acaba de empezar y, por ello, es muy probable que acaben resolviendo los problemas que hemos presentado en estas notas, pero, por el momento, tienen razón Bender *et al*. y Chomsky *et al.* al sostener que estas IIAA son radicalmente diferentes a la competencia lingüística de un ser humano.

Pienso que aciertan plenamente Bender *et al.* cuando sostienen, contrariamente a Piantadosi, que ChatGPT no tiene semántica alguna y que el significado y la coherencia no está en el texto que genera la IA, sino en la mente del lector humano. La impresión de dialogar con la máquina es un puro espejismo, como lo sería entrar en conversación con un loro, que sí tiene cierta capacidad mental, o con una muñeca parlante, que no tiene ninguna. Recuérdense a este propósito aquellos tamagotchis que provocaron algún suicidio infantil. Buena parte del "milagro" de ChatGPT está en nuestra cabeza, que atribuye a la IA estados mentales análogos a los nuestros. Los humanos somos una especie fuertemente simbólica; podemos atribuir necesidades y sentimientos a objetos inanimados. Somos un niño que quiere auxiliar a un oso de peluche tirado en la basura: el oso no sufre; sufre solamente el niño.

Bender, E. M., Gebru, T., McMillan-Major, A. y Shmitchell, S. «On the dangers of stochastic parrots: Can language models be too big?» en *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 2021, 610–623.

Chomsky, N. (1986): *Knowledge of Language. Its nature, Origin and Use.* New York, Praeger.

Chomsky, N., Roberts, I. y Watumull, J. "The false promise of ChatGPT" en *The New York Times.* 2023. Disponible en: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Katzir, R. "Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi". 2023. Disponible en: https://lingbuzz.net/lingbuzz/007190.

Mendívil-Giró, J.L. "Un loro estocástico en la habitación china. ¿Qué nos enseña CHATGPT sobre la mente humana?" en *Letras libres*, 2023, nº 262, 6-22. Disponible en:

https://letraslibres.com/revista/un-loro-estocastico-en-la-habitacion-china-que-nos-ensena-chatgpt-sobre-la-mente-humana/01/07/2023/

Pintadosi, S. “Modern language models refute Chomsky’s approach to language”. 2023. Disponible en: https://lingbuzz.net/lingbuzz/007180.