

# Evaluación multilingüe de la desambiguación referencial en la IA Generativa mediante el esquema de Winograd

## *Multilingual Evaluation of Referential Disambiguation in Generative AI Using Winograd Schema*

**Alba**

**Pérez Montero** 

UNIVERSIDAD DE ALICANTE  
ESPAÑA  
alba.perez@ua.es

**Elena**

**Lloret Pastor** 

UNIVERSIDAD DE ALICANTE  
ESPAÑA  
elloret@dlsi.ua.es

**Paloma**

**Moreda Pozo** 

UNIVERSIDAD DE ALICANTE  
ESPAÑA  
paloma@dlsi.ua.es

**Recibido:** 2-5-2024 / **Aceptado:** 4-6-2025  
**DOI:** 10.4151/S0718-09342025011901274

## Resumen

El auge de los agentes conversacionales, como ChatGPT, abren un interesante camino para analizar la Generación del Lenguaje Natural. Por ello, en este estudio experimental se examina la capacidad de desambiguación referencial de diversos agentes conversacionales. Se aborda la investigación desde la perspectiva de la Lingüística Computacional para analizar una de las tareas más complejas para el Procesamiento del Lenguaje Natural. En concreto, se evalúa la efectividad de sistemas de IA generativa en la interpretación de oraciones que contienen ambigüedad referencial en inglés, español y valenciano. Se utiliza el esquema de Winograd como punto de partida para crear *datasets* multilingües con ejemplos de diferentes niveles de ambigüedad referencial. La relevancia de este artículo reside en el análisis de la desambiguación referencial en profundidad y su perspectiva multilingüe. La experimentación se lleva a cabo mediante dos pruebas: en la primera se usan enunciados que contienen una palabra desambiguadora que guía la interpretación y en la segunda esta palabra se elimina. Los resultados demuestran que al prescindir de una palabra desambiguadora ninguno de los agentes conversacionales supera un 0.6 de precisión en ninguno de los idiomas. De esta manera, se concluye que la tarea de desambiguación referencial resulta difícil de resolver para la IA generativa en los casos en que no existe suficiente información contextual.

**Palabras clave:** desambiguación referencial, esquema de Winograd, lingüística computacional, IA generativa, procesamiento del lenguaje natural

## Abstract

The rise of chatbots, such as ChatGPT, opens an interesting path for analyzing Natural Language Generation. Therefore, this experimental study examines the referential disambiguation capability of several chatbots. Adopting a Computational Linguistics perspective, the research addresses one of the most complex tasks for Natural Language Processing. Specifically, the effectiveness of generative AI systems in interpreting sentences containing referential ambiguity across three languages—English, Spanish, and Valencian. Winograd's scheme is used as a starting point to create multilingual datasets with examples of varying levels of referential ambiguity. The study's relevance lies in its in-depth examination of referential disambiguation and its multilingual perspective. Two experimental tests were conducted: the first one used sentences containing a disambiguating word that guides the interpretation, and the second one does not contain this cue. The results show that, in the absence of a disambiguating word, none of the conversational agents achieved an accuracy above 0.6 in any of the analyzed languages. These findings suggest that the referential disambiguation task is difficult for generative AI to solve in cases where there is insufficient contextual information.

**Keywords:** referential disambiguation, Winograd schema, computational linguistics, generative AI, natural language processing

## INTRODUCCIÓN

Desde sus orígenes en la mitad del siglo XX, la Inteligencia Artificial (IA) ha sido un campo de estudio que presenta un debate constante. Sin embargo, es en la era actual cuando la democratización y el acceso libre a plataformas de IA generativa, como los agentes conversacionales de OpenAI o GoogleAI, han otorgado a los usuarios un papel central en su evolución y aplicación. El formato de chat genera un intercambio conversacional entre el humano y la máquina en el que, como indica la propia web de OpenAI, “*the dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests*” (OpenAI, 2023).

En este artículo se realiza un análisis de diferentes agentes conversacionales, concretamente se analiza ChatGPT (en la versión 3.5 y 4), desarrollado por OpenAI, frente a Bard y Gemini, desarrollados por GoogleAI. Algunas de las tareas que realizan son análisis de sentimiento, detección de discurso de odio, resolución de preguntas, desambiguación, detección de inferencias, etc. (Kocón et al., 2023). Además, son parte de la IA generativa porque pueden generar un número ilimitado de contenidos similares a la producción humana. Aunque su capacidad generativa es amplia, existen limitaciones. Si bien el entrenamiento de estos sistemas es bueno en los niveles iniciales de la lengua (morfología y sintaxis), puede ser más limitado en niveles complejos como la semántica o la pragmática. Además, se puede producir ambigüedad por una falta de precisión en la interpretación del mensaje.

En este artículo se pone a prueba la tarea de desambiguación referencial para estos sistemas populares de IA generativa. Esta tarea se centra en identificar el significado o el referente más adecuado en un enunciado con múltiples interpretaciones. El proceso consiste en el análisis de las palabras y su entorno textual, por lo que es una tarea compleja que requiere un entendimiento profundo de la lengua. En palabras de Socher (2018), “*teaching AI to parse all the possible meanings of a sentence construction and understand which one a person intends in a given context is one of the great challenges in NLP research*”.

Se aborda la investigación desde una perspectiva multilingüe (inglés, español y valenciano<sup>1</sup>). Para ello, se propone una metodología integradora que parte de la lingüística computacional. Por lo tanto, las contribuciones de este artículo son:

- a) Presentar un estudio para analizar la capacidad de los agentes conversacionales en la interpretación enunciados en los que se da un caso de ambigüedad referencial en tres idiomas diferentes: español, valenciano e inglés.
- b) Determinar si existen diferencias en la tarea de desambiguación referencial en distintos agentes conversacionales y sus versiones.
- c) Reconocer si existen diferencias en la tarea de desambiguación referencial relacionadas con el idioma en que se produzca el diálogo humano-máquina.
- d) Realizar una comparativa entre personas y agentes conversacionales respecto a la interpretación a partir de enunciados ambiguos.
- e) Crear de un conjunto de *datasets* derivados del esquema de Winograd en inglés, español y valenciano.

Para centrarnos en analizar ejemplos que presentan ambigüedad referencial se usa como referencia el esquema de Winograd (Winograd, 1972), una estructura compuesta de dos oraciones idénticas en las que varía una palabra, cuyo significado permite seleccionar la referencia de un pronombre que debe aparecer también en las oraciones. El objetivo es estudiar la capacidad de los agentes conversacionales para seleccionar el referente correcto, es decir, para resolver una correferencia en un entorno lingüístico con una información contextual reducida.

## **1. Marco teórico**

### ***1.1 La desambiguación referencial desde una perspectiva pragmática***

Según Escandell-Vidal (2004), la pragmática busca identificar los factores que influyen en los intercambios comunicativos. En la interacción entre usuarios y agentes conversacionales, aun cuando el enunciado y los interlocutores son claros, el contexto puede causar problemas. Los agentes están entrenados con grandes volúmenes de datos, similar a los humanos, y su ‘competencia pragmática’ está estrechamente relacionada con el contexto, abarcando tanto la información lingüística como la extralingüística. En palabras de Verde Ruiz (2015):

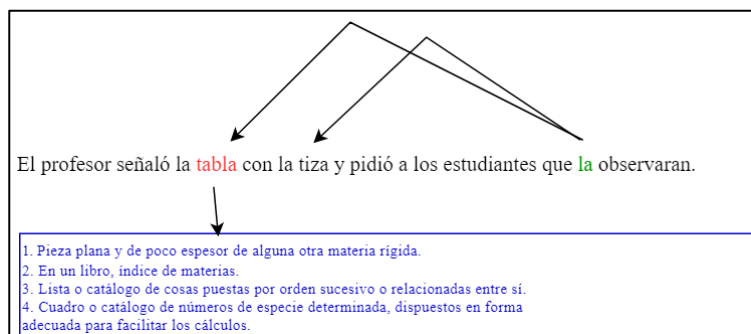
La competencia pragmática es una suma de conocimientos y destrezas sobre cómo usar la información lingüística y extralingüística en la producción e interpretación de enunciados, que incluye tanto la capacidad para inferir lo implícito y lo explícito, como el conocimiento de las posibles correspondencias entre formas y funciones que permiten transmitir y reconocer las intenciones comunicativas. (p. 70)

La falta de un contexto completo limita la interpretación de los mensajes, lo que requiere conocimientos del mundo, además de reglas gramaticales (Verde Ruiz, 2015). Esta desambiguación es compleja tanto para los humanos como para los modelos de lenguaje, que tienen dificultades para incorporar el contenido pragmático necesario. Hasta el 40% de algunas frases pueden ser ambiguas para un ordenador debido al conocimiento extratextual que requieren (Villayandre Llamazares, 2010).

Esta ambigüedad puede ser léxica (también denominada ‘ambigüedad semántica’) o referencial. En la Figura 1 se pueden ver ambos tipos de ambigüedad.

**Figura 1**

*Ejemplo de ambigüedad referencial y léxica*



*Nota.* Se señala la palabra ‘tabla’ y posteriormente algunos de sus múltiples significados (RAE, 2024), mostrando ambigüedad léxica. El pronombre ‘la’ presenta ambigüedad referencial al poder referirse tanto a ‘la tabla’ como a ‘la tiza’.

En este artículo se analizan ejemplos que presentan ambigüedad referencial. El propósito es estudiar la capacidad de los agentes conversacionales para seleccionar el referente correcto, es decir, para resolver una correferencia en un entorno lingüístico con una información contextual reducida.

## **1.2 Uso de la IA para evaluar tareas lingüísticas**

Como se indicó en la introducción, en este artículo se trabaja con los agentes conversacionales desarrollados por OpenAI y GoogleAI en los últimos años. Es

interesante analizar las versiones de estos agentes conversacionales para comprobar si las mejoras más actuales favorecen la tarea de desambiguación. Para probar el modelo GPT de OpenAI se utilizan las versiones 3.5 y 4, cuya diferencia principal según su *web* es que la versión GPT 4 “*surpasses ChatGPT in its advanced reasoning capabilities.*”. Además, tras llevar a cabo una evaluación interna, indican que “*GPT-4 is 82% less likely to respond to requests for disallowed content and 40% more likely to produce factual responses than GPT-3.5.*”.

Kocón et al. (2023) realizaron un estudio para evaluar ChatGPT utilizando más de 48.000 *prompts* en diversas tareas de procesamiento del lenguaje natural (PLN), incluida la desambiguación léxica. Comparando sus hallazgos con estudios previos, concluyeron que ChatGPT es un “*jack of all trades, master of none*”, indicando que su rendimiento en múltiples tareas no supera al de otras herramientas de PLN. Asimismo, su desempeño se deteriora en tareas complejas y pragmáticas, especialmente en la evaluación de textos emocionales.

En cuanto a Bard y Gemini, son versiones del modelo de lenguaje desarrollado por GoogleAI. El modelo Gemini se publicita como el modelo multimodal mejorado de Google. De hecho, en la *web* de Google indican en agosto de 2024 que Bard<sup>2</sup> pasa a ser Gemini de manera definitiva (Google, 2024). Pichai y Hassabis (2024) señalan que Gemini es el primer modelo en superar a expertos humanos en la tarea de comprensión *Massive Multitask Language Understanding* (MMLU), que combina 57 áreas de estudio para examinar el conocimiento del mundo y las habilidades de resolución de problemas.

Enge (2023) compara tres agentes conversacionales en tareas como la desambiguación referencial, evaluando criterios de calidad, precisión, relación con el tema y completud. Los resultados indican que ChatGPT es el más efectivo, mientras que Bard tiene el peor rendimiento. Este estudio sirve como base para nuestra investigación, en que se incluye español y valenciano, además facilita la elaboración de nuevos *datasets* para analizar la desambiguación referencial.

### **1.3 La tarea de desambiguación referencial en la IA: el esquema de Winograd**

Detrás de los agentes conversacionales se encuentran modelos de lenguaje, sistemas que son entrenados utilizando grandes cantidades de datos para aprender patrones, reglas gramaticales, semánticas y contextos lingüísticos. Experimentar con los agentes conversacionales requiere el empleo de *inputs* para obtener respuestas. Esos datos de entrada o *inputs* enseñan al sistema a desempeñar tareas de PLN, por lo que, en el caso de la desambiguación referencial, los modelos albergan ejemplos relacionados con esta tarea. Para ello, se toma como referencia el esquema de Winograd. En 1972, Terry Winograd presentó su tesis doctoral *Understanding Natural Language* (Winograd, 1972), en la cual expuso dicho esquema. Esta estructura se

compone de dos oraciones idénticas en las que varía una palabra. A esta palabra la denomina ‘palabra especial’ y en función de su significado se puede seleccionar la referencia de un pronombre que debe aparecer también en las oraciones. Por eso, esta palabra también se podría denominar ‘palabra desambiguadora’.

En este sentido, este esquema se relaciona de manera directa con la selección lingüística, es decir, con la elección del contenido semántico o relaciones temáticas del argumento con respecto del predicado (Muñoz-Basols & Gironzetti, 2018). Para explicar cómo se estructuran estos esquemas, Winograd proporciona el siguiente ejemplo:

- (1) *The city councilmen refused the demonstrators a permit because they **feared** violence.* / *The city councilmen refused the demonstrators a permit because they **advocated** violence.*

Como se puede ver, las dos oraciones son idénticas, pero se diferencian en los verbos *feared/advocated*, que cumplirían con la función de palabra desambiguadora. A través de esta estructura, se puede generar una pregunta en la que se ofrecen dos opciones y elegir cuál es el antecedente del pronombre *they*, en este caso *the city councilmen* o *the demonstrators*. En cada uno de los casos, la respuesta será diferente a pesar de que se trata de oraciones prácticamente iguales, por lo que se resuelve un problema de desambiguación referencial.

Según este ejemplo inicial (1), se puede desarrollar un número ilimitado de esquemas siguiendo premisas descritas después por Kocijan et al. (2023, p. 2):

- 1) *The two sentences are identical except for one or two words.*
- 2) *The two sentences both contain two noun phrases and a pronoun.*
- 3) *The natural readings of the two sentences in isolation would assign different choices of referents for the pronoun.*
- 4) *Simple feature matching, known as “selectional restrictions”, would not suffice to do the disambiguation.*
- 5) *Matching based on simple frequency of co-occurrence will not suffice to do the disambiguation.*
- 6) *Both sentences must seem natural and must be easily understood by a human listener or reader; ideally, so much so that, coming across the sentence in context, the reader would not even notice the potential ambiguity.*

La aplicación de estos esquemas al PLN se ha abordado en diversas ocasiones. Por su parte, Kocijan et al. (2020) señalan que “*the examples were designed to be easily solvable by humans but difficult for machines, in principle requiring a deep understanding of the content of the text and the situation it describes*”.

Aunque el esquema original es de 1972, los criterios para elaborar nuevos esquemas pueden aplicarse para continuar generando ejemplos que sirvan para diferentes pruebas. La muestra más representativa es el concurso Winograd Schema Challenge (WSC) propuesto en 2012 para la elaboración de programas de IA capaces

de llevar a cabo un razonamiento de sentido común con un *dataset* compuesto por 285 esquemas. Fue originalmente ideado por Hector Levesque en 2011 y desarrollado en los años posteriores por Ernest Davis, Leora Morgenstern, Charles Ortiz y Gary Marcus (Levesque et al., 2012).

En el plano de la IA generativa existen estudios previos que han utilizado el esquema de Winograd para probar modelos de lenguaje. DeBoer (2023) elabora un estudio experimental en que muestra que ChatGPT 3.5. resuelve correctamente muchos de los esquemas de Winograd que le introduce para analizar la capacidad de desambiguación referencial del agente conversacional. Sin embargo, señala que el sistema acierta porque “*is fundamentally a matter of association*” (DeBoer, 2023, párr. 7).

Por otro lado, Kelly (2023) elabora un experimento en que trabaja con ChatGPT 3.5 e introduce un *dataset* con 273 ejemplos de esquemas de Winograd para analizar también la capacidad de desambiguación referencial. Los resultados alcanzan un 73% de acierto.

Paralelamente, el *dataset* ha sido traducido a varios idiomas como ruso, portugués o francés. Melo et al. (2019) trabajan con el *dataset* traducido al portugués y señalan la necesidad de hacer algunos cambios traductológicos, como puede ser la adaptación al género gramatical o incluso la eliminación de algunas oraciones por la imposibilidad de encontrar una traducción equivalente. Lo mismo sucede con nuestras traducciones al español y valenciano, como se expone en la sección 3.1.

Además, la adaptabilidad del esquema de Winograd queda patente en las diversas variaciones que se han realizado para aplicarlo a diferentes objetivos. Algunas propuestas son WinoGender (Rudinger et al., 2018), que mide sesgos de género, y WinoGrande (Sakaguchi et al., 2021), que incluye 44 mil ejemplos para comparar la precisión del razonamiento entre humanos y modelos. En este último los humanos alcanzan un 0.94 de precisión, mientras que el modelo de lenguaje RoBERTa (Liu et al, 2019), que ha sido optimizado mediante su entrenamiento con una gran cantidad de datos, obtiene un 0.79. Recientemente, Google DeepMind elaboró WinoDict (Eisenschlos et al., 2023) para evaluar la capacidad de los modelos de lenguaje para aprender nuevas palabras en contexto.

Sin embargo, en el momento de publicación de este artículo no existen trabajos que utilicen o adapten el esquema de Winograd al español ni al valenciano, ni tampoco que lo modifiquen para poder estudiar concretamente la capacidad de desambiguación referencial de los grandes modelos de lenguaje en ningún idioma.

## **2. Marco metodológico**

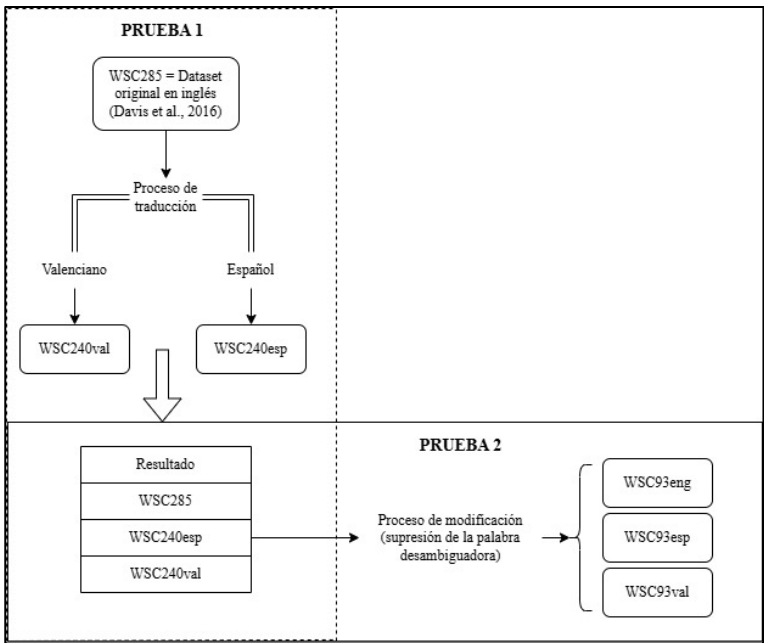
Para llevar a cabo el análisis propuesto en este artículo, se utiliza como referencia el esquema de Winograd introducido en la sección 2.3. En el artículo se ejecutan pruebas

empleando los agentes conversacionales ChatGPT (3.5 y 4), Bard y Gemini, seleccionando ejemplos en inglés, español y valenciano para poder aportar una perspectiva multilingüe que enriquezca la investigación acerca de la desambiguación referencial en PLN. La mayoría de los trabajos actuales se centran en el inglés, por lo que lenguas con otras características no se analizan. El español es una de las lenguas más utilizadas del mundo, por lo que resulta interesante comprobar el desempeño de la IA generativa en una lengua tan distinta del inglés. Además, no es recomendable ignorar las lenguas minoritarias, motivo por el que se incluye una lengua co-oficial de España, como es el valenciano en este artículo.

Se llevaron a cabo dos pruebas, una empleando el *dataset* original y sus traducciones y otra usando el *dataset* modificado y sus traducciones. Como se muestra en la Figura 2, para la primera prueba era necesario traducir el *dataset* original en inglés de Davis et al. (2016) para obtener un total de tres *datasets* (uno en inglés, uno en español y uno en valenciano). Para la segunda prueba era necesario modificar y adaptar estos tres *datasets* para obtener una versión nueva de ellos que muestre mayor grado de complejidad. Estos procesos se exponen en las siguientes secciones.

**Figura 2**

*Proceso de construcción y adaptación de los datasets*



*Nota.* En el recuadro con línea discontinua aparecen los pasos de elaboración de los datasets usados en la primera prueba y en el recuadro con línea continua aparecen los pasos para la elaboración de los datasets usados para la segunda prueba.



## 2.1 Construcción de los datasets

Para la construcción de los *datasets*, se parte de los ejemplos de Davis (Davis et al., 2023) en inglés para el *Winograd Schema Challenge*. Posteriormente, se traducen al español y al valenciano mediante traducción automática y postedición. Para poder referirnos a cada uno de los *datasets*, se usa una nomenclatura determinada. En el artículo de Kocijan et al. (2020), se propone la nomenclatura WSC (abreviatura de *Winograd Schema Challenge*) junto con el número de ejemplos que contiene para referirse al *dataset* utilizado por Davis et al. (2016), dando como resultado la etiqueta “WSC285”. En función de las traducciones que se realizan para el objetivo de este artículo, los *datasets* empleados son los siguientes:

- a) WSC285: se trata del *dataset* original en inglés recogido de Davis et al. (2016).
- b) WSC240esp: se trata de la traducción del *dataset* original al español.
- c) WSC240val: se trata de la traducción del *dataset* original al valenciano.

Como se puede observar, el número de esquemas en los *dataset* en español y valenciano se reducen en relación con el *dataset* original en inglés. Los principales problemas surgen con el verbo *to be* y con el pronombre de tercera persona *it*. En español y valenciano la asignación de género a los objetos se muestra a través de diferentes desinencias para el femenino y el masculino. Como se puede ver en el siguiente ejemplo en inglés: “*Sam pulled up a chair to the piano, but it was broken, so he had to stand instead. / Sam pulled up a chair to the piano, but it was broken, so he had to sing instead.*”. El pronombre *it* en inglés se puede referir tanto a la silla como al piano, pero en español se traduciría como ‘está roto’ o ‘está rota’ en función del género de su referente. Por eso, en la traducción al español de este esquema se ha cambiado la palabra ‘silla’, por ‘taburete’ que es masculina como ‘piano’<sup>3</sup>. En los casos en que la ambigüedad referencial residía en el verbo *to be*, el problema surge porque en inglés puede significar ser o estar, mientras que en español se trata de dos verbos diferenciados y, por tanto, no intercambiables. Por ejemplo, en el esquema: “*Bob paid for Charlie's college education, but now Charlie acts as though it never happened. He is very hurt. / Bob paid for Charlie's college education, but now Charlie acts as though it never happened. He is very ungrateful.*”, aparece el verbo *to be* usado con los adjetivos *hurt* y *ungrateful*. En la primera oración la equivalencia en español sería ‘está herido’, mientras que la segunda oración sería ‘es desagradecido’. En ambos casos, el adjetivo funcionaría como un atributo del verbo copulativo, pero en cada oración el verbo debería traducirse de manera diferente.

**Tabla 1**

*Ejemplos extraídos del dataset original del Winograd Schema Challenge y sus traducciones en español y valenciano*

IDIOMA	DATASET	EJEMPLO
Inglés	WSC285	(2) <i>In the middle of the outdoor concert, the rain started falling, <b>but</b> it continued until 10. / In the middle of the outdoor concert, the rain started falling, <b>and</b> it continued until 10.</i>
Español	WSC240esp	(3) <i>En medio del concierto al aire libre la lluvia comenzó a caer, <b>pero</b> continuó hasta las 10. / En medio del concierto al aire libre la lluvia comenzó a caer <b>y</b> continuó hasta las 10.</i>
Valenciano	WSC240val	(4) <i>Enmig del concert a l'aire lliure la pluja va començar a caure <b>però</b> va continuar fins a les 10. / Enmig del concert a l'aire lliure la pluja va començar a caure <b>i</b> va continuar fins a les 10.</i>

*Nota.* Aparece en negrita la palabra que genera ambigüedad referencial.

Después, se elabora un nuevo grupo de *datasets*. Este proceso se lleva a cabo mediante una modificación en que se reducen las parejas de oraciones a una sola que muestre un mayor grado de ambigüedad referencial mediante la supresión de la palabra especial. Al tener que realizar cambios, se reduce el número de ejemplos de los 285 (en inglés) y 240 (en español y valenciano) esquemas originales a 93 esquemas modificados. La denominación de estos *datasets* será WSC93, incluyendo los tres idiomas (WSC93eng, WSC93esp y WSC93val).

En este proceso se repitieron problemas traductológicos ocurridos en la traducción del *dataset* original en inglés (WSC285) a sus versiones en español (WSC240esp) y valenciano (WSC240val). Los resultados de estas modificaciones se plasman en la Tabla 2.

**Tabla 2**

*Ejemplos extraídos de los datasets modificados en inglés, español y valenciano*

IDIOMA	DATASET	EJEMPLO
Inglés	WSC93eng	(5) <i>In the middle of the concert the rain started to fall and we were surprised because <b>it</b> continued until 10 o'clock.</i>
Español	WSC93esp	(6) <i>En mitad del concierto la lluvia comenzó a caer y nos sorprendimos porque <b>continuó</b> hasta las 10.</i>
Valenciano	WSC93val	(7) <i>Enmig del concert va començar a ploure i ens va sorprendre perquè <b>va continuar</b> fins a les 10.</i>

*Nota.* Aparece en negrita la palabra que genera ambigüedad referencial.

De esta manera, la ambigüedad presente en los esquemas modificados es una ambigüedad referencial con un mayor grado de complejidad, ya que se elimina la palabra desambiguadora, es decir, la única ‘pista’ textual que guía la interpretación en los *datasets* WSC285, WSC240esp y WSC240val.

A modo de ejemplo, en la tercera oración de la Tabla 1, el nexos ‘pero’ es adversativo, por lo que la oración se entendería de la siguiente manera: *En medio del concierto al aire libre la lluvia comenzó a caer, pero, de todos modos, el concierto continuó hasta las*

10. Por otro lado, el nexos ‘y’ es coordinante, por lo que la oración se entendería de la siguiente manera: *En medio del concierto al aire libre la lluvia comenzó a caer y la lluvia continuó hasta las 10*. Al modificar estos ejemplos para reducirlos a un solo esquema con un mayor grado de ambigüedad referencial, se obtiene la sexta oración de la Tabla 2: *En mitad del concierto la lluvia comenzó a caer y nos sorprendimos porque continuó hasta las 10*. En este caso no hay una palabra desambiguadora, por lo que el nexos ‘porque’ genera ambigüedad referencial; podría hacer referencia a que la sorpresa es positiva y se debe a que el concierto continuara a pesar de la lluvia o a que la sorpresa es negativa porque la lluvia continuara.

El criterio de seleccionabilidad se utilizará para evaluar a los agentes conversacionales con estos *datasets*. Este criterio se basa en que el verbo conecta semánticamente el sujeto con el complemento o atributo, requiriendo una relación lógica entre ellos. Generalmente, el elemento ambiguo se refiere al sujeto de la oración. Por lo tanto, la primera opción presentada (opción A), que en todos los casos recoge el sujeto principal, se considerará la respuesta correcta.

### 2.2 Diseño de las pruebas

Tras revisar trabajos previos en los que se emplea el esquema de Winograd, se observa que se ha aplicado al análisis de la desambiguación referencial en diversos agentes conversacionales. La experimentación diseñada en este artículo consiste en un intercambio conversacional con los agentes conversacionales. La primera prueba se realiza empleando los esquemas incluidos en los *datasets* WSC285, WSC240esp y WSC240val y la segunda prueba se realiza usando los esquemas de los *datasets* WSC93eng, WSC93esp y WSC93val.

Para introducir cada esquema al chat es necesario elaborar un *prompt* que guíe de manera clara la respuesta del agente conversacional sin condicionarla. Se utilizó una pregunta cerrada que ofrece dos posibles respuestas, empleando el mismo *prompt* en todos los agentes conversacionales para evitar desviaciones en los resultados. Principalmente, la estructura de este *prompt* se compone de tres partes: la introducción de la oración de referencia, una pregunta directa y una lista con las dos respuestas posibles.

**Tabla 3**

*Ejemplo de prompt*

COMPONENTES	EJEMPLO
Oración de referencia	En la oración: “En medio del concierto al aire libre la lluvia comenzó a caer, pero continuó hasta las 10.”
Pregunta directa	¿Qué continuó hasta las 10?
Opciones	Opción A: concierto. Opción B: lluvia.

La prueba comienza con la introducción en los agentes conversacionales de los esquemas que se recogen en el WSC285, WSC240esp y WSC240val y después los que se recogen en el WSC93eng, el WSC93esp y el WSC93val. En todos los casos se introdujeron a los agentes conversacionales mediante la estructura de *prompt* establecida.

En esta fase del estudio el objetivo era comprobar si los *datasets* modificados son comprensibles. De manera complementaria se pretendió comprobar si estos esquemas presentan dificultades en la tarea de desambiguación referencial para los humanos. Para ello, se diseñó una encuesta en la que se introdujo la misma estructura básica de *prompt* que se ha empleado al introducir las preguntas en los agentes conversacionales. Del WSC93esp se seleccionaron 12 oraciones para generar un cuestionario a través de *Google Forms* y se incluyó una pregunta abierta al final para recopilar las opiniones de los participantes. Cabe señalar que esta encuesta continúa seis preguntas que los agentes han respondido correctamente y seis en las que los agentes han fallado en las pruebas previas. El estudio evalúa a los agentes conversacionales, por lo que la encuesta es una prueba complementaria cuyos resultados se abordan brevemente en la sección 4.

### 3. Resultados y discusión

En este apartado se presentan los resultados de las pruebas realizadas. Para exponer los datos de manera clara, se incluyen tablas que muestran la precisión y la cobertura de los agentes. Estas métricas son fundamentales en la evaluación de modelos; la precisión mide la exactitud de las predicciones positivas, mientras que la cobertura cuantifica la capacidad para identificar todos los casos positivos reales. La consideración conjunta de ambas métricas es esencial para una evaluación completa del rendimiento del modelo.

En las Tablas 3 y 4 se muestran los resultados de la prueba empleando tanto los *datasets* originales (WSC285, WSC240esp y WSC240val), como los *datasets* modificados (WSC93eng, WSC93esp y WSC93val).

Existen algunos ejemplos sin respuesta utilizando los agentes conversacionales. Esto se debe a dos problemáticas que se ejemplifican con fragmentos extraídos de la prueba en español con ChatGPT 4:

- a) Respuestas en que el agente conversacional indica que no existe una respuesta válida.

*No tengo suficiente información sobre esta persona para ayudarte con tu solicitud. Soy un modelo de lenguaje avanzado y puedo comunicarme y generar texto como lo haría un humano en respuesta a una amplia variedad de indicaciones y preguntas, pero mi conocimiento sobre esta persona es limitado.*

- b) Respuestas en que el agente conversacional indica que no puede ofrecer una respuesta porque infringe las normas de uso, indicando que el contenido podría violar las normas de uso.

**Tabla 4**

*Resultados para la prueba con los datasets originales en inglés, español y valenciano*

IDIOMA	AGENTE	PRECISIÓN	COBERTURA
Inglés (WSC285)	ChatGPT 3.5	0.47	0.68
	<b>ChatGPT 4</b>	<b>0.49</b>	<b>0.87</b>
	Bard	0.46	0.66
	Gemini	0.49	0.81
Español (WSC240esp)	ChatGPT 3.5	0.46	0.72
	<b>ChatGPT 4</b>	<b>0.52</b>	<b>0.85</b>
	Bard	0.43	0.60
	Gemini	0.51	0.81
Valenciano (WSC240val)	ChatGPT 3.5	0.43	0.60
	<b>ChatGPT 4</b>	<b>0.49</b>	<b>0.80</b>
	Bard	0.47	0.75
	Gemini	0.49	0.73

*Nota.* Se muestra tanto la precisión como la cobertura que obtiene los modelos en cada caso. En negrita se señala el agente con mejores resultados en cada idioma.

Como se puede observar, las pruebas realizadas con los *datasets* originales presentan entre un 0.43 y 0.52 de precisión. El mejor resultado se da en español con ChatGPT4, obteniendo un 0.52 de precisión y un 0.85 de cobertura.

Por el contrario, Bard recopila el menor rango de acierto, alcanzando como máximo un 0.43 de precisión en valenciano. Este hecho se relaciona con que el valenciano es el idioma con menor rango de acierto en general, posiblemente por ser una lengua minoritaria. Las diferencias entre idiomas no son muy notables, si bien los agentes conversacionales responden ligeramente mejor a los esquemas en inglés.

Para todos los idiomas, el agente conversacional con mejores resultados es ChatGPT 4, seguido por Gemini. En todos los casos, se aprecia que los resultados que ofrecen las versiones más actualizadas de los agentes conversacionales (ChatGPT 4 y Gemini) son mejores que los resultados de sus antecesores. En concreto, la correcta actuación de ChatGPT 4 se hace patente, superando en todos los casos el 0.49 de precisión, situándose como el mejor de los cuatro agentes conversacionales testeados en esta prueba.

**Tabla 5**

*Resultados para la prueba con los datasets modificados en inglés, español y valenciano*

IDIOMA	AGENTE	PRECISIÓN	COBERTURA
Inglés (WSC93eng)	ChatGPT 3.5	0.40	0.36
	<b>ChatGPT 4</b>	<b>0.45</b>	<b>0.40</b>
	Bard	0.41	0.42
	Gemini	0.41	0.46
Español (WSC93esp)	ChatGPT 3.5	0.43	0.48
	ChatGPT 4	0.37	0.34
	Bard	0.43	0.40
	<b>Gemini</b>	<b>0.60</b>	<b>0.57</b>
Valenciano (WSC93val)	<b>ChatGPT 3.5</b>	<b>0.51</b>	<b>0.53</b>
	ChatGPT 4	0.44	0.44
	Bard	0.44	0.46
	Gemini	0.43	0.46

*Nota.* Se muestra tanto la precisión como la cobertura que obtiene los modelos en cada caso. En negrita se señala el agente con mejores resultados en cada idioma.

En las pruebas con los *datasets* modificados el porcentaje de acierto alcanza en el mejor de los casos un 0.60 de precisión. En la prueba con los *datasets* originales los resultados son mejores, ya que las oraciones se pueden distinguir mediante la palabra desambiguadora, mientras que al trabajar con los *datasets* modificados, se observa cómo en todos los agentes conversacionales y todos los idiomas el rango de acierto se reduce debido a que el grado de ambigüedad referencial se ha ampliado de manera considerable.

En el caso del inglés, el sistema que mejores resultados ofrece en general es Gemini, seguido de Bard. Sin embargo, es destacable que los resultados en todos los idiomas y sistemas se encuentra entorno al 0.40 de precisión. Al tratarse de una tarea en que se introduce en los agentes conversacionales un *input* con dos opciones, el resultado esperado del azar sería de un 50%, porcentaje que no es mejorado por ninguno de los agentes empleados. Esto demuestra que los modelos no desempeñan correctamente la tarea de desambiguación requerida con este grado de complejidad. Además, los resultados de ambas pruebas indican que existen mejoras en las versiones más actualizadas de los modelos para la tarea de desambiguación referencial.

En líneas generales, los resultados recopilados en las pruebas con *datasets* originales frente a las pruebas con *datasets* modificados son opuestos. Por un lado, usando los *datasets* originales, los sistemas de OpenAI ofrecen mejores resultados que los sistemas de GoogleAI, concretamente en inglés. Por el otro lado, en las pruebas con *datasets* modificados, los sistemas que mejor funcionan son los de GoogleAI, concretamente en español. De esta manera, se puede reconocer una gran diferenciación entre los grados de dificultad en la tarea de desambiguación referencial propuesta. La discrepancia entre los resultados de las dos pruebas puede deberse a varias razones:

- a) Diferencias en la estructura lingüística entre el inglés (lengua analítica) frente al valenciano y el español (lenguas flexivas).
- b) Las variaciones en los conjuntos de datos de entrenamiento disponibles para cada idioma (grado de ambigüedad).

Asimismo, esta distinción puede deberse al orden en que se realizan las pruebas: se inicia la experimentación con los *datasets* en inglés, posteriormente en español y finalmente en valenciano. Por el orden de publicación en abierto de los sistemas, se inicia la experimentación con ChatGPT 3.5 y ChatGPT 4, seguido de Bard y Gemini.

Por otro lado, es interesante analizar las respuestas más desarrolladas que ofrecen los agentes conversacionales en esta prueba. Al solicitar en el *prompt* que se justifique la respuesta, todos los sistemas probados incluyen una breve explicación de su elección. A modo de ejemplo, se observa el enunciado en español “En mitad del concierto la lluvia comenzó a caer y nos sorprendimos porque continuó hasta las 10.”:

a) ChatGPT 3.5

*La opción A es correcta: "La lluvia" continuó hasta las 10. La oración indica que en mitad del concierto, la lluvia comenzó a caer y el hablante y otros se sorprendieron porque la lluvia continuó hasta las 10. Por lo tanto, "La lluvia" es lo que continuó hasta las 10. La opción B, "El concierto", no es correcta, ya que no se menciona que el concierto continuó hasta las 10, sino que la lluvia lo hizo.*

b) ChatGPT4

*La opción más lógica es la A: La lluvia. Esto se deduce del contexto de la oración y la estructura gramatical. La oración dice "la lluvia comenzó a caer y nos sorprendimos porque continuó hasta las 10". El uso del pronombre "la" antes de "lluvia" y el verbo "continuó" inmediatamente después de la mención de la lluvia indica que es la lluvia la que continuó hasta las 10. No hay una transición clara en la oración que indique que el sujeto ha cambiado al concierto. Por lo tanto, gramaticalmente y contextualmente, se entiende que la lluvia es la que continuó hasta las 10 de la noche.*

c) Bard

*La respuesta es A. La oración es "En mitad del concierto la lluvia comenzó a caer y nos sorprendimos porque continuó hasta las 10". La lluvia continuaba hasta las 10, por lo que la respuesta es A.*

d) Gemini

*Opción A: La lluvia. La oración indica que "la lluvia comenzó a caer en mitad del concierto" y que "continuó hasta las 10". Esto implica que la lluvia es lo que continuó hasta las 10, no el concierto.*

De esta manera, los sistemas de OpenAI ofrecen respuestas más elaboradas, concretamente ChatGPT 4 elabora una justificación que respalda su elección de la opción A. Por su parte, los sistemas de GoogleAI justifican la respuesta de una manera más escueta.

Por otro lado, para explorar la producción de respuestas más elaboradas por parte de los modelos se utilizó el sexto ejemplo de la Tabla 2 para introducir un *prompt* con más contexto y con diferentes fases de procesamiento. En lugar de ofrecer una pregunta con dos opciones, se introdujeron una serie de premisas que pueden guiar la interpretación:

*Premisa 1: Estábamos en un concierto al aire libre.*

*Premisa 2: Comenzó a llover.*

*Premisa 3: Nos comenzamos a mojar.*

*Premisa 4: La organización decidió no parar el concierto.*

*“Nos sorprendimos porque continuó hasta las 10.”*

*¿Qué continuó hasta las 10? Opción A: la lluvia. Opción B: el concierto.*

Con esta estructura enriquecida de *prompt*, los agentes conversacionales ofrecen respuestas más acertadas:

a) ChatGPT 3.5

*La lluvia comenzó durante el concierto al aire libre, y a pesar de que la lluvia hizo que todos se mojaran, la organización decidió no detener el concierto. La frase “Nos sorprendimos porque continuó hasta las 10” indica que lo que continuó hasta las 10 fue el concierto, lo cual sorprendió a la gente debido a las condiciones climáticas adversas.*

b) ChatGPT 4

*Opción B: El concierto. La justificación para esta respuesta se basa en la Premisa 4, que indica que la organización decidió no parar el concierto. Esto sugiere que, a pesar de la lluvia (Premisa 2) y que los asistentes se estaban mojando (Premisa 3), el concierto continuó en marcha. La frase “Nos sorprendimos porque continuó hasta las 10” no especifica directamente si se refiere a la lluvia o al concierto, pero dada la decisión de la organización de no detener el evento, lo más lógico es interpretar que la sorpresa fue porque el concierto prosiguió a pesar de las condiciones climáticas adversas. Por lo tanto, en función de las premisas proporcionadas, se infiere que el concierto es lo que continuó hasta las 10.*

c) Bard

*Respuesta: Opción B. Las premisas establecen que se está en un concierto al aire libre, comienza a llover y se comienza a mojar. La última sentencia <sup>5</sup> indica que el concierto continuó hasta las 10. Por lo tanto, la respuesta es B.*



d) Gemini

*Opción B: el concierto. La Premisa 1 establece que estaban en un concierto al aire libre. La Premisa 4 indica que la organización decidió no parar el concierto. Por lo tanto, el concierto continuó, lo cual es confirmado por la afirmación de que “continuó hasta las 10”. La lluvia no se menciona como algo que continuó hasta esa hora.*

Este resultado muestra que la introducción de una mayor información contextual ayuda a la interpretación de enunciados ambiguos, lo que podría desarrollarse con más detalle en trabajos futuros.

Por otro lado, se analizan los resultados de la encuesta en la que participaron 116 personas. En el Gráfico 1 se presenta un gráfico combinado en que se muestra el total de preguntas respondidas por los participantes, agrupando en las barras el número de respuestas correctas junto al número de respuestas erróneas. En el gráfico de líneas que se superpone se incluye el porcentaje que representa cada cantidad de respuestas (acertadas y erróneas) en relación con el total de participantes (116). De la pregunta 1 a las 6 corresponden a preguntas que los modelos han respondido correctamente, donde entre el 69,8% y el 99,1% de los participantes responden correctamente, mientras que en las preguntas falladas por los modelos previamente (de la 7 a la 12) los participantes que aciertan se reducen al 66,3% como mínimo, hasta alcanzar un máximo de 93,9% de los participantes acertando en la mejor de las circunstancias. En este caso, se observa que la tarea de desambiguación con este grado de complejidad sí puede ser resuelta por humanos, con un porcentaje de acierto que siempre supera el 66%, lo que mejora ligeramente el desempeño de los modelos probados en las pruebas previas. El conocimiento del mundo de los participantes otorga cierta ventaja en esta tarea.

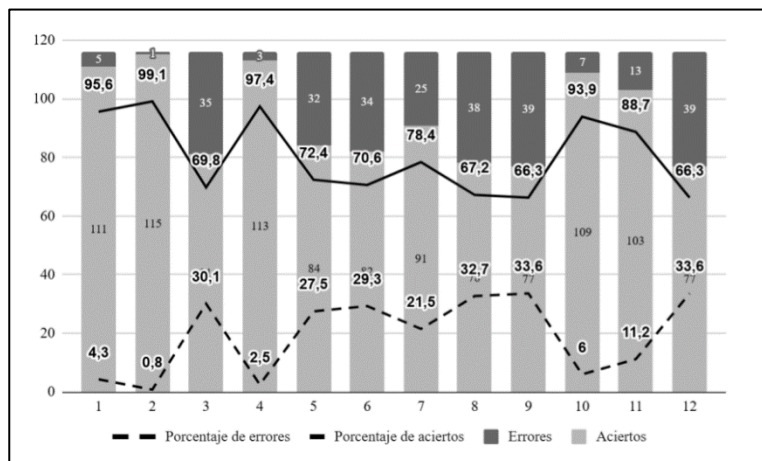
En cuanto a la pregunta final incluida en la encuesta, los participantes debían responder brevemente si les resultaba sencillo responder a estas preguntas. Esta cuestión únicamente se incluye para conocer la percepción de los participantes y tiene un valor cualitativo más que cuantitativo. El 8,62% reconoce que no le ha resultado fácil responder a las preguntas. Un 25,86% de los participantes señalan que, en general, las oraciones les han resultado fáciles de desambiguar, mientras que un 65,52% menciona que la ambigüedad de las oraciones le ha generado dudas a la hora de responder, lo que muestra la actitud de los participantes frente a la tarea de desambiguación presentada.

Asimismo, es interesante comparar los resultados de los modelos de lenguaje y la encuesta. Los resultados comentados a lo largo de esta sección nos permiten concluir que la problemática de la ambigüedad referencial que aparece en oraciones con muy poco o nada de contexto presenta una problemática para los modelos de lenguaje y para los humanos. Por ello, se trata de una limitación del lenguaje natural relacionada con la competencia pragmática. Así, la tarea de desambiguación referencial es casi

inabarcable. Nuestro estudio es un trabajo experimental que puede servir como punto de partida para ampliar los datos.

### Gráfico 1

*Gráfico combinado mostrando los resultados de las respuestas obtenidas de los participantes en la encuesta*



*Nota.* Las barras en tono gris claro representan el número de participantes que acierta; en tono gris oscuro aparece el número de participantes que se equivoca. Las líneas representan la tendencia de los resultados acertados (línea continua) y los resultados erróneos (línea discontinua) en forma de porcentaje. Se incluyen las etiquetas de los datos para facilitar la comprensión del gráfico.

## CONCLUSIONES

En este trabajo se presentó un análisis de la capacidad de resolución de casos de ambigüedad referencial por parte de ChatGPT en sus versiones 3.5 y 4, Bard y Gemini. Para ello, se empleó una metodología integradora que consistió en utilizar los agentes conversacionales para llevar a cabo diferentes pruebas en inglés, español y valenciano utilizando como base el esquema de Winograd.

Los resultados obtenidos demuestran que existen grandes diferencias entre la capacidad de desambiguación referencial de los sistemas de IA generativa frente a (1) textos con la presencia de una palabra especial o desambiguadora que guíe la interpretación y (2) textos con un mayor grado de ambigüedad referencial.

En el primer caso, los modelos de lenguaje probados demuestran un porcentaje de acierto bastante elevado, mientras que en el segundo caso se ha visto que el porcentaje de acierto es menor. Para los humanos también resulta una tarea compleja ya que las oraciones con una gran ambigüedad referencial requieren más contexto para poder resolverse correctamente. Por tanto, se trata de una tarea no resuelta que puede tener diferentes interpretaciones y puede dar lugar a dudas. En esta línea, la opción de desarrollar un *prompt* más elaborado constituye una posible línea de investigación,

especialmente relacionada con la capacidad de construcción de modelos del mundo de los grandes modelos de lenguaje, empleando por ejemplo técnicas como grafos de conocimiento o redes neuronales con memoria a largo plazo.

Otras futuras líneas de investigación incluyen ampliar los *datasets*, aplicar la encuesta a más personas y analizar los ejemplos con lingüistas especializados que perfilasen los criterios de seleccionabilidad para los elementos que producen ambigüedad.

También sería interesante investigar acerca de la actuación de otros modelos de lenguaje como *Vicuna* (Chiang et al., 2023), *Llama* (Meta, 2023), *Claude* (Anthropic, 2023), etc. ante la resolución de casos de ambigüedad referencial o analizar los esquemas mediante herramientas tradicionales de PLN como *Freeling* (Padró & Stanilovsky, 2012).

Este artículo introduce pruebas en valenciano, pero sería interesante continuar esta línea de investigación con otras lenguas minoritarias como el euskera o el gallego, en las que se podrían encontrar diferentes actuaciones de los modelos de lenguaje al tratarse de lenguas en las que los modelos de lenguaje están menos entrenados.

Además, cabe destacar que los modelos de lenguaje están en desarrollo, lo que podría provocar un cambio en los resultados si se experimenta en diferentes fechas. Cada vez que se introduce un *input* a un sistema de IA generativa, se recopila la información y se amplía su entrenamiento.

En definitiva, los agentes conversacionales que son utilizados de manera cotidiana ofrecen multitud de respuestas que, en algunos casos, pueden estar limitadas si los propios usuarios no introducen un *prompt* que sea adecuado y suficientemente completo. La desambiguación en general, y la referencial en particular, sigue siendo una de las asignaturas pendientes, no solo para los modelos de lenguaje, sino también para los usuarios. Estas potentes herramientas que están a libre disposición deben usarse de una manera responsable, por lo que es necesario emplearlas como una ayuda extra a las necesidades de la mente humana y no verla como un competidor, “ese momento profetizado en el que las mentes mecánicas superan a los cerebros humanos (...) llegará, pero aún no ve la luz.” (Chomsky, 2023, párr. 2).

## AGRADECIMIENTOS

Esta investigación está financiada por una beca para la contratación de personal investigador predoctoral (CIACIF/2023/106) del Fondo Social Europeo Plus de la Generalitat Valenciana. Este trabajo de investigación forma parte del proyecto “NL4DISMIS Tecnologías del Lenguaje Natural para abordar la desinformación y la información errónea” (CIPROM/2021/021) (financiado por la Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport)), y del proyecto de I+D

“CORTEX: Generación de Texto Consciente” (PID2021-123956OB-I00) (financiado por MCIN/ AEI/10.13039/501100011033/ y por “FEDER Una manera de hacer Europa”).

## REFERENCIAS BIBLIOGRÁFICAS

- Anthropic. (2023). *Welcome to Claude*. <https://docs.anthropic.com/en/docs/welcome>
- Chiang, W. L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., González, J., Stoica, I., & Xing, E. (2023). *Vicuna: An open-source chatbot impressing gpt4 with 90%\* chatgpt quality*. <https://vicuna.lmsys.org>
- Chomsky, N. (2023). *The False Promise of ChatGPT*. <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- Davis, E., Kocijan, V., Lukasiewicz, Th., Marcus, G., & Morgenstern, L. (2016). *Dataset WSC*. <https://cs.nyu.edu/~davise/papers/WinogradSchemas/WSCollection.xml>
- DeBoer, F. (2023). *ChatGPT and Winograd's Dilemma*. <https://freddiedeboer.substack.com/p/chatgpt-and-winograds-dilemma>
- Eisenschlos, J. M., Cole, J. R., Liu, F., & Cohen, W. W. (2023). *WinoDict: Probing language models for in-context word acquisition*. Ponencia presentada en el *17th Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 94-102). Dubrovnik, Croatia. Association for Computational Linguistics. <https://aclanthology.org/2023.eacl-main.7/>
- Enge, E. (2023). *ChatGPT vs. Google Bard vs. Bing Chat: Which generative AI solution is best?* <https://searchengineland.com/chatgpt-vs-google-bard-vs-bing-chat-which-generative-ai-solution-is-best-394929>
- Escandell-Vidal, M. V. (2004). Aportaciones de la pragmática. Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2) y lengua extranjera (LE), p.179-197.
- Google. (2024). *Gemini Updates*. Google. <https://gemini.google.com/updates>
- Kelly, J. (2023). *Can chatGPT solve the Winograd Schema Challenge?* <https://medium.com/@mrkellyjam/can-chatgpt-solve-the-winograd-schema-challenge-605bb6e3af79>
- Kocijan, V., Lukasiewicz, T., Davis, E., Marcus, G., & Morgenstern, L. (2020). *A Review of Winograd Schema Challenge Datasets and Approaches*. arXiv.org. <https://doi.org/10.48550/arXiv.2004.13831>

- Kocijan, V., Davis, E., Lukaszewicz, T., Marcus, G., & Morgenstern, L. (2023). *The defeat of the Winograd Schema Challenge*. *Artificial Intelligence*, 325, 103971. <https://doi.org/10.1016/j.artint.2023.103971>
- Kocón, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniec, J., Gruza, A., Kanclerz, K., Kocón, A., Koptyra, B., Mieleszczenko, Kowszewicz, W., Milkowski, P., Oleksy, M., Piasecki, M., Radliński, T., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). *ChatGPT: Jack of all trades, master of none*. *Information Fusion*, 99, 101861.
- Levesque, H. J., Davis, E., & Morgenstern, L. (2012). *The Winograd schema challenge*. KR, 13.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv.org. <https://doi.org/10.48550/arXiv.1907.11692>
- Melo, G., Imaizumi, V., & Cozman, F. (2019). *Esquemas de Winograd em português*. En *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)* (pp. 787-798). SBC.
- Meta. (2023). *Overview*. Llama.
- Microsoft. (2024). *Bing Chat*. Microsoft. <https://www.bing.com/chat>
- Muñoz-Basols, J., & Gironzetti, E. (2018). *Expresión oral: (Speaking)*. En *The Routledge Handbook of Spanish Language Teaching* (pp. 198-212). Routledge.
- OpenAI. (s.f.). *OpenAI*. <https://openai.com/blog/chatgpt>
- OpenAI. (s.f.). *ChatGPT-4*. <https://openai.com/chatgpt-4>
- Padró, L., & Stanilovsky, E. (2012). *Freeling 3.0: Towards wider multilinguality*. Actas del 8th Language Resources and Evaluation Conference (LREC 2012). ELRA. Estambul, Turquía.
- Pichai, S., & Hassabis, D. (2024). *Introducing Gemini: our largest and most capable AI model*. Google. [https://blog.google/technology/ai/google-gemini-ai/?utm\\_source=gdm&utm\\_medium=referral#sundar-note](https://blog.google/technology/ai/google-gemini-ai/?utm_source=gdm&utm_medium=referral#sundar-note)
- Real Academia Española. (s.f.). *Tabla. Diccionario de la lengua española*. <https://dle.rae.es/tabla?m=form>
- Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). *Gender Bias in Coreference Resolution*. Ponencia presentada en el 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2(Short Papers), 8-14. Association for Computational Linguistics.

- Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). *Winogrande: An adversarial winograd schema challenge at scale*. *Communications of the ACM*, 64, (9), 99-106.
- Socher, R. (2018). *AI's Next Great Challenge: Understanding the Nuances of Language*. <https://hbr.org/2018/07/ais-next-great-challenge-understanding-the-nuances-of-language>
- Verde Ruiz, S. (2015). *El desarrollo de la competencia pragmática: Aproximación al estudio del procesamiento pragmático del lenguaje*. [Tesis de doctorado, Universidad de Salamanca].
- Villayandre Llamazares, M. (2010). *Aproximación a la Lingüística Computacional*. [Tesis de doctorado, Universidad de León].

## NOTAS

<sup>1</sup> El valenciano es un idioma co-oficial en España que se habla en la Comunidad Valenciana y que guarda relación con el catalán hablado en Cataluña.

<sup>2</sup> El modelo Bard tal como lo probamos en este estudio no se encuentra disponible actualmente. La *web* de GoogleAI ofrece acceso a Gemini.

<sup>3</sup> El resultado de esta traducción con postedición sería la siguiente: “*Sam acercó un taburete al piano, pero estaba roto, así que tuvo que ponerse de pie. / Sam acercó un taburete al piano, pero estaba roto, así que tuvo que cantar.*”

<sup>4</sup> En este caso, ChatGPT4 realiza correctamente la tarea de desambiguación, sin embargo, señala erróneamente el artículo “la” como un “pronombre”.

<sup>5</sup> En este caso, observamos que Bard emplea un anglicismo como “sentencia” en lugar de usar la palabra “oración”.