¿Qué tan bien entienden los LLM los Modismos Colombianos?

Santiago Bobadilla (s.bobadilla@uniandes.edu.co), Juan Diego Osorio (jd.osorioc1@uniandes.edu.co)
María Alejandra Pinzón (ma.pinzonr1@uniandes.edu.co), Ignacio Chaparro (i.chaparro@uniandes.edu.co)

I. Planteamiento del problema

Desde el lanzamiento de ChatGPT en noviembre de 2022, se ha observado una rápida adopción y un aumento significativo en el uso de LLM. Entre los patrones encontrados en el estudio *"The widespread adoption of large language model-assisted writing across society"*, se observa que aproximadamente el 24% de los comunicados de prensa corporativos utilizaban escritura asistida por LLM para finales de 2024. Incluso, se encontró que, para la misma fecha, los comunicados de prensa de las Naciones Unidas (ONU) alcanzaban cerca del 14%. [1]

En principio, el estudio ve dichos patrones como positivos. Las compañías pueden producir comunicados más rápido y de manera más rentable, especialmente en áreas que requieren actualizaciones frecuentes y/o difusión de información compleja. No obstante, se advierte que la creciente dependencia de contenido asistido por LLM puede introducir desafíos al divulgar mensajes con pocos detalles o genéricos, volviendo la información menos creíble y generando desconfianza pública en la autenticidad de estos. [1]

Esta credibilidad se vuelve particularmente peligrosa al ver las limitaciones que tienen los LLM para comprender el lenguaje humano de una manera significativa, especialmente el lenguaje figurado, como los modismos. Nicole Kobie, en su artículo "*What idioms teach us about AI and its evolution*" expone un ejemplo contundente de este riesgo: imaginemos que un chatbot (similar a la línea de teléfono NHS 111) de atención médica interpreta "feeling blue" de manera literal. Esto podría llevar a una valoración errada, en lugar de dirigir a la persona a recursos de salud mental. [2] Si pensamos ahora en un comunicado de prensa generado a partir de un discurso y/o entrevista en el cual se hace uso de modismos, ¿qué credibilidad tendría la información presentada si la interpretación del modismo es errada?

El artículo, basado en gran medida en la investigación de Tom Pickard de la Universidad de Sheffield, sostiene que el problema principal reside en que los LLM no entienden el lenguaje. Sino que por el contrario "destroza" el texto, reduciendo palabras y letras a tokens (que están "desprovistos de cualquier significado") en vectores y números multidimensionales que se usan posteriormente para entrenamiento. [2]

Si observamos este problema y los riesgos implicados, junto con el estudio "*Es igual pero no es lo mismo: ¿Distinguen los LLM las variedades del español?*", podemos ver que los LLM están entrenados predominantemente para identificar el lenguaje del español peninsular, mientras que la comprensión de las variantes dialectales y modismos regionales es, por defecto, un área de vulnerabilidad. Esto es notorio al ver que la puntuación promedio de los nueve modelos evaluados en el estudio (incluyendo GPT-4o, GPT-4o-mini, y varios modelos Llama y Gemma) para la variedad peninsular es de 20.2/30; mientras que el promedio de las puntuaciones obtenidas por los mismos modelos para la variedad de español andino y caribeño es de aproximadamente 8.2/30 y 9.8/30 respectivamente. [3]

Por eso, con el fin de evaluar el entendimiento de los distintos dialectos colombianos por parte de los LLM el presente estudio busca cuantificar el entendimiento de los modismos presente en todo el territorio colombiano (español andino y caribeño). Para eso, se busca construir una amplia base de conocimiento que incluya el modismo, su definición y un ejemplo, con base en el Sistema Gestor Lexicográfico (DICOL) del Instituto Caro y Cuervo [4], y la cuarta edición del diccionario de Colombianismos de la Academia Colombiana de la Lengua. [5]

Cada modismo será evaluado en definición y contextos de uso, para cada uno de los LLM. Se busca ver I) la definición dada por cada LLM y su similitud (por coseno) con la definición original, II) la capacidad de reconocer (clasificación binaria) si el ejemplo de nuestra base de conocimiento es un modismo o no y III) la capacidad de remplazar el modismo por un sinónimo con significado similar, teniendo como contexto solo el ejemplo proporcionado.

II. Descripción y análisis de los datos

El conjunto de datos presentado se compone a partir de dos fuentes lexicográficas colombianas: el Sistema Gestor Lexicográfico (DICOL, *Diccionario de Colombianismos*) del Instituto Caro y Cuervo [5] y el Diccionario de Colombianismos de la Academia Colombiana de la Lengua en su cuarta edición (BDC, *Breve Diccionario de Colombianismos*) [4].

A priori al análisis de los datos, vale la pena explorar cada base de datos a profundidad. La primera fuente que empleamos, DICOL (Diccionario de Colombianismos, 2018), constituye el resultado de un proyecto colaborativo iniciado en 2010 entre el Instituto Caro y Cuervo y la Academia Colombiana de la Lengua, desarrollado formalmente entre 2015 y 2017. [6] Esta obra contiene 540 páginas, fue dirigida por Carmen Millán como directora del Instituto Caro y Cuervo, coordinada por Nancy Rozo Melo (coordinadora del Instituto Caro y Cuervo) y María Clara Henríquez Guarín (coordinadora de la Academia Colombiana de la Lengua), con un equipo de investigadores

integrado principalmente por egresados de la Maestría en Lingüística Hispánica del Instituto Caro y Cuervo y becarios de la Escuela de Lexicografía Hispánica de la Asociación de Academias de la Lengua Española, entre otros especialistas. [7] El diccionario comprende cerca de 8000 definiciones, 6000 entradas, 1500 expresiones y 4500 ejemplos. [8] DICOL, está disponible públicamente a través de la página web oficial, lo que permite el acceso libre para consultas académica. Acorde con la misión del Instituto Caro: "contribuir al conocimiento, la promoción y la difusión de las lenguas y culturas de Colombia". [9]

La segunda fuente que empleamos, BDC (4ª edición revisada, 2012), parte de una proyecto tradicional lexicográfico iniciado por la Academia Colombiana de la Lengua en 1975 con su primera edición, revisada, hasta llegar a la cuarta edición en 2012 con aproximadamente 126 páginas. El diccionario fue dirigido por Jaime Posada como director de la Academia Colombiana de la Lengua y coordinado por Carlos Patiño Roselli. Este estudio muestra que el conjunto de datos se apoya en una fuente con más de 47 años de evolución metodológica y respaldo institucional sólido. [5] Además, el diccionario incluye un sistema de marcación con: abreviaturas gramaticales (adj., adv., f., fr., etc.), marcas sociolingüísticas (coloq., despect., obsol., etc.) y marcas regionales que abarcan los 32 departamentos colombianos. El BDC, como obra protegida por derechos de autor de la Academia Colombiana de la Lengua, está disponible para uso académico bajo el principio de la cita y uso legítimo establecido en la Ley 23 de 1982 de Colombia sobre derechos de autor. [5]

El conjunto de datos resultante (base de conocimiento creada para este estudio) se organizó con base en la siguiente estructura de tres columnas: palabra, significado y ejemplo; donde, cada fila o registro corresponde a una expresión completa, ya sea una palabra individual o una frase idiomática, acompañada de la definición y una oración que ilustra como se emplea en el habla cotidiana. En total, el conjunto de datos generados para este estudio integra 10,195 registros, incluyendo tanto modismos de una sola palabra, como expresiones compuestas. Del total, 7,505 provienen del Breve Diccionario de Colombianismos de la Academia Colombiana de la Lengua (BDC) y 2,690 del Diccionario de Colombianismos (DICOL) del Instituto Caro y Cuervo (véase Figura 1).

La primera columna referente a la variable "palabra" corresponde al modismo o expresión idiomática registrado en formato de texto estandarizado sin alterar su sentido lingüístico. La variable "significado" contiene la definición asociada al modismo, que describe su sentido figurado o uso particular dentro del español colombiano. La variable "ejemplo" contiene una oración que ilustra el ejemplo real del modismo bajo la definición dada. Todos los modismos se encuentran en formato de texto plano y se modelaron en un archivo CSV disponible en GitHub[1]

Para la construcción del conjunto de datos del Diccionario de Colombianismos, se interactuó con la página oficial Instituto Caro y Cuervo [4] para extraer la totalidad de las entradas. La fuente de datos original, base de datos documental, presentaba una notable desorganización y múltiples tipos de registros inconsistentes entre sí. Esto requirió la creación de una serie de reglas de procesamiento para extraer de manera fiable todas las definiciones y ejemplos registrados por los autores. Ya con los datos extraídos de manera correcta, se organizaron bajo el formato estructurado ya definido para este estudio. Se debe notar que, en los casos donde una palabra tenía más de un significado, se generaron registros individuales para cada definición: manteniendo la idiomática original (ya sea solo el modismo, o su uso compuesto) con sus respectivas definiciones y ejemplos. Finalmente, se realizó una limpieza de estos datos ya estructurados eliminaron las entradas que carecían de significado (tras verificar su ausencia en la fuente original). Este método de expansión de registros es la razón por la cual existe un número considerable de entradas que, si bien tienen una definición, carecen de un ejemplo asociado, ya que no todos los significados en la fuente original contaban con su respectivo ejemplo.

En el caso del BDC, el proceso de preprocesamiento para el conjunto de datos que presentamos se centró en conservar la riqueza léxica y contextual propia de la fuente. Para esto tuvimos en cuenta diferentes decisiones de modelar los datos, dada la estructura vista en la fuente original. [4] En primer lugar, cuando un término presentaba más de una definición, cada una se trató como un registro distinto dentro de nuestra base de conocimiento, permitiendo representar las variaciones semánticas de un mismo modismo. Asimismo, se preservaron los modismos formados por varias palabras, respetando su estructura original. En cuanto a los modismos que se encontraban en formas flexionadas, se normalizaron los lemas a su forma base, por ejemplo "cansado, da" se registró como "cansado", esto con el fin de unificar el vocabulario.

Teniendo en cuenta las condiciones de ambas fuentes, se determinó la cantidad de expresiones que no tenían ningún valor en la columna "ejemplo", es decir, modismos que cuentan únicamente con su definición. Para estos casos la propuesta que se realizará posteriormente consiste en emplear herramientas de IA generativa como ChatGPT o Gemini con la búsqueda en la web, para complementar los ejemplos de uso a partir del modismo y significado que obtuvimos de la fuente original. Este procedimiento se plantea como un proceso semiautomático con validación humana, en el cual los ejemplos generados serían posteriormente revisados y ajustados para garantizar su adecuación lingüística y su correspondencia con el español colombiano.

Finalmente, este conjunto es adecuado para el problema planteado pues reúne expresiones idiomáticas del dialecto colombiano (andino y caribe) a partir de fuentes seleccionadas provenientes de instituciones con trayectoria reconocida en investigación lexicográfica. Permitiéndonos tener datos de confianza para evaluar los distintos modelos.

Ya con este conjunto final estructurado, procedemos a realizar un análisis descriptivo sobre los datos, para entender un poco las características presentes de estos. Este análisis se presenta por medio de las siguientes gráficas:

[1] https://github.com/NLP-UAndes/Proyecto

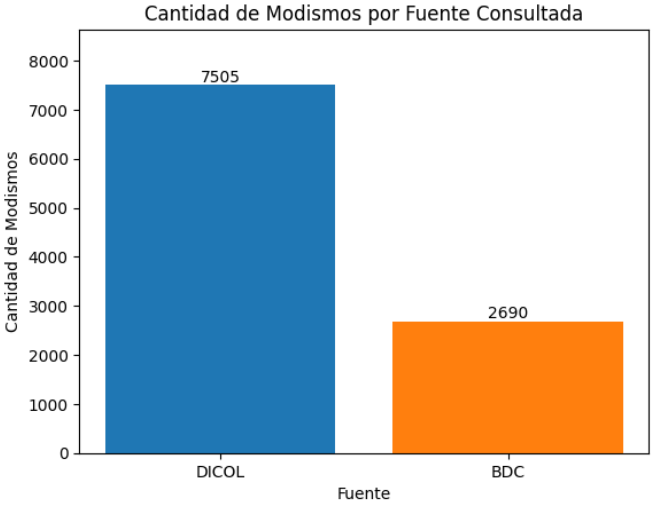

Figura 1. Distribución de registros por fuente lexicográfica

La figura 1 ilustra la contribución cuantitativa de cada fuente al conjunto de datos consolidado. Se observa que el Diccionario de Colombianismos (DICOL) es la fuente principal, aportando 7,505 modismos, lo que representa la gran mayoría de los registros. Por su parte, el Breve Diccionario de Colombianismos (BDC) complementa el corpus con 2,690 entradas. Esta distribución resalta la predominancia de la fuente DICOL en la compilación final, que consolida un total de 10,195 registros de modismos colombianos. Si bien observamos que varias palabras estaban en ambas bases de datos, dado que no tenían el mismo significado, para esos casos mantuvimos las dos ocurrencias.

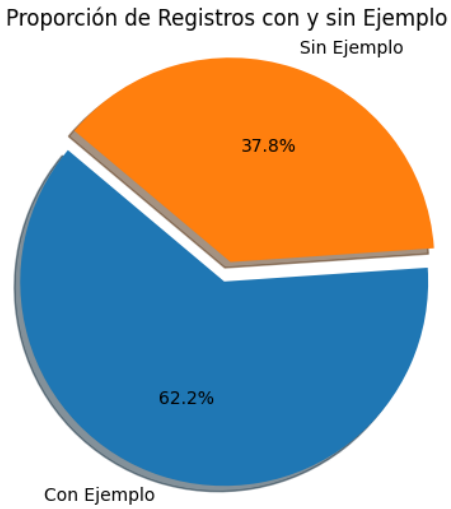

Figura 2. Proporción de registros con y sin ejemplo de uso.

La figura 2 muestra la completitud del conjunto de datos en lo que respecta a los ejemplos de uso. Se evidencia que una mayoría del 62.2% de los registros cuenta con un ejemplo asociado, lo cual proporciona un valioso contexto. Sin embargo, una proporción considerable del 37.8% de las entradas carece de esta información. Esta ausencia representa una oportunidad de mejora clave para el enriquecimiento del corpus. Por ello, se proyecta como trabajo futuro el uso de un LLM para generar automáticamente ejemplos pertinentes para estas entradas. Esto será clave para mejorar los resultados del proyecto

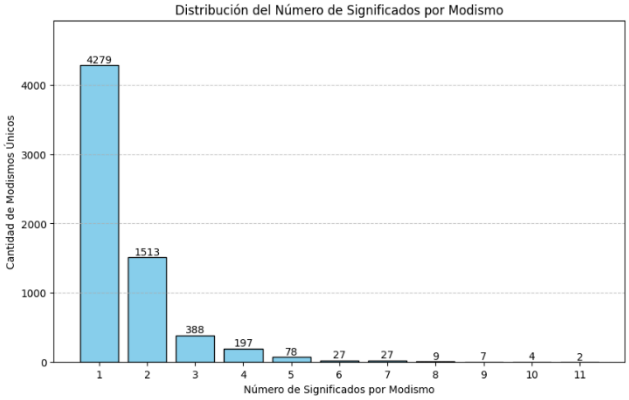

Figura 3. Distribución de la cantidad de definiciones por modismo.

En la figura 3, está el conteo palabras con base a la cantidad de significados que tiene presente. Podemos ver que la mayoría de nuestros datos se centran en una cantidad mínima de significados (1), mostrando que la mayoría de nuestros modismos son muy particulares y tiene usos muy puntuales. Por otro lado, tenemos una mínima cantidad de modismos con una gran cantidad de definiciones, en donde resaltan modismos como “mano”, que tiene múltiples usos y definiciones en la cultura colombiana. Esto va a implicar un reto para los modelos: por un lado, deben ser capaces de entender términos muy particulares y de baja frecuencia de uso y por otro, deben gestionar la ambigüedad de los modismos con una gran flexibilidad semántica.

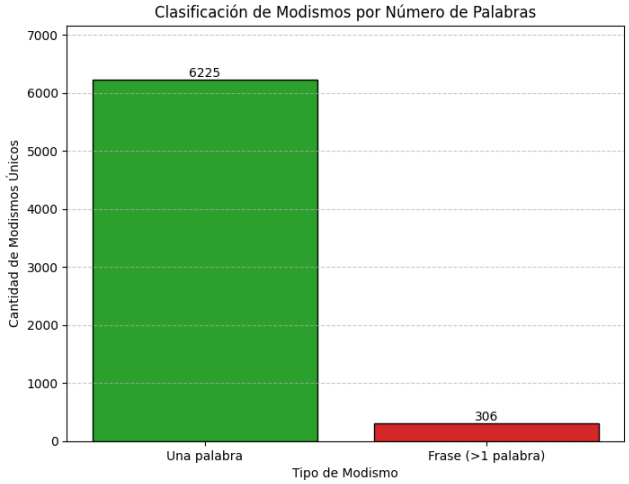

Figura 4.

En la figura 4, al clasificar los modismos por su estructura, podemos ver que la gran mayoría de nuestros datos, con 6,225 entradas, corresponden a modismos de una sola palabra. Esto muestra que una parte fundamental del reto se centra en el entendimiento de términos individuales que adquieren un significado especial en el contexto colombiano. Por otro lado, encontramos una cantidad mucho menor, de 306 modismos, que son frases compuestas por más de una palabra. En estos modismos compuestos resaltan principalmente las frases coloquialmente conocidas en Colombia, como por ejemplo “al que no quiere caldo, se le dan dos tazas”. Estas dos estructuras van a implicar un reto interesante para los modelos, ya que no solo deben capturar el significado no literal de palabras únicas, sino también aprender a tratar secuencias enteras de palabras como una única unidad semántica

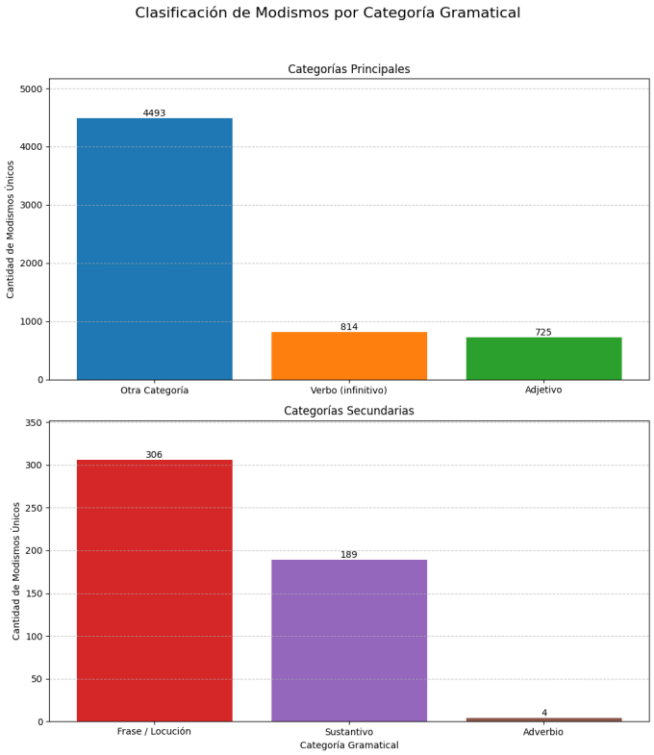

Figura 5.

En la figura 5, al realizar una clasificación gramatical de los modismos mediante reglas heurísticas, podemos ver una distribución muy particular. La categoría más grande es "Otra Categoría" con 4,493 entradas, lo que indica que una gran cantidad de modismos no siguen las terminaciones morfológicas estándar de sustantivos, adjetivos o verbos, presentando una forma irregular.

Le siguen en importancia los verbos en infinitivo (814) y los adjetivos (725), mostrando una fuerte presencia de acciones y cualidades en el corpus. Estas irregularidades creemos van a causar un deterioro en los modelos ya que no podrán apoyarse consistentemente en la morfología de la palabra para inferir su función.

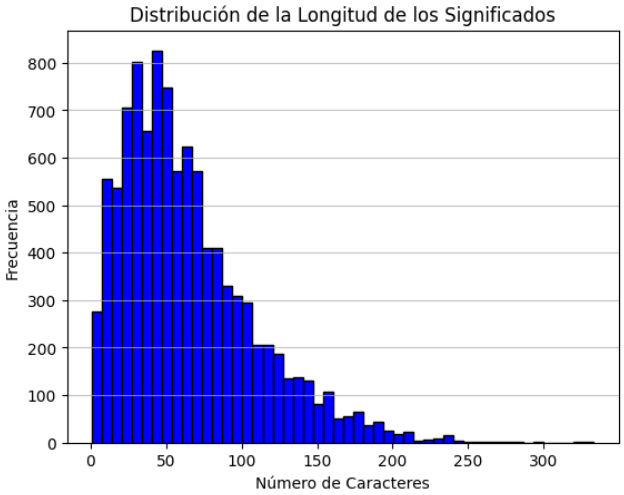

Figura 6. Histograma que muestra la distribución de frecuencias de la longitud de los significados en el dataset.

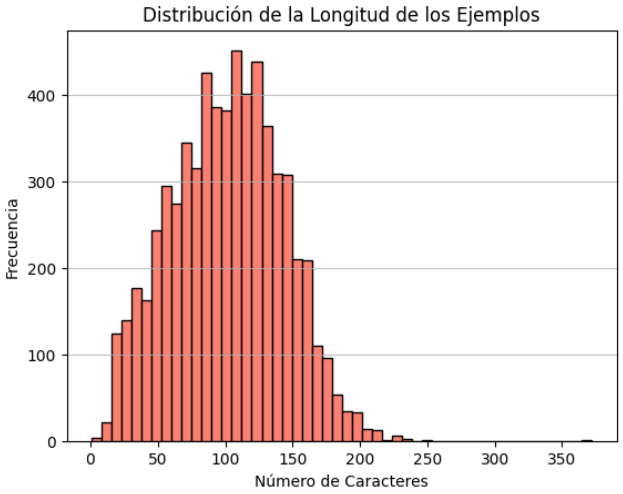

Figura 7. Histograma que muestra la distribución de frecuencias de la longitud de los ejemplos en el dataset.

Finalmente, se generaron histogramas que muestran la distribución de la longitud de los significados y los ejemplos. El histograma de los significados confirma que las definiciones son predominantemente cortas y directas, con una alta frecuencia en el rango de 25 a 50 caracteres y con altos casos en los que simplemente se agrega un sinónimo como definición. Esta brevedad tiene implicaciones técnicas importantes; por ejemplo, al momento de vectorizar el texto, sugiere que una estrategia de concatenación de las representaciones individuales de los tokens podría ser más efectiva, preservando así la información de cada palabra.

Por otro lado, el histograma de los ejemplos muestra que estos son más extensos, con una longitud más frecuente en el intervalo de 100 a 125 caracteres. Este hallazgo es crucial para el futuro proceso semiautomático de generación de ejemplos con un LLM, ya que indica que sería valioso solicitar explícitamente que los nuevos ejemplos se generen con longitudes dentro de un intervalo de confianza alrededor de este pico de frecuencia, asegurando así la consistencia y calidad del corpus enriquecido.

III. REVISIÓN DE LITERATURA TÉCNICA

En el estado del arte revisado no se encontraron estudios puntuales respecto a la evaluación de modismos colombianos haciendo uso de LLM. En sí, de los estudios encontrados y que se describirán a continuación, la mayoría se basan en métricas y resultados obtenidos primordialmente sobre el idioma inglés. [10] [11] [12]

Dado ese marco, y a priori a los baselines encontrados sobre el reconocimiento de modismos por parte de diferentes LLM, partimos de dos conceptos que consideramos claves: Primero, el sesgo lingüístico de los LLM por los datos de entrenamiento [13] y segundo, las limitantes de los modelos de representación de palabras (embeddings) para representar modismos. [10] Puntos sobre los cuales se soporta todo LLM actual.

El primer concepto se detalla en el artículo "*El Sesgo Lingüístico Digital (SLD) en la inteligencia artificial:*

implicaciones para los modelos de lenguaje masivos en español." [13] (Anexo[2] 1), donde introducen el término: Sesgo Lingüístico Digital (SLD), sobre el idioma español. El término describe la hibridez lingüística que generan los LLM en dos niveles: I) Interlingüístico: En relación con la base del inglés utilizada frecuentemente para entrenar los modelos. II) Intralingüístico: En relación con las distintas variedades geográficas y sociales del español.

El principal punto es que los LLM actuales no reflejan adecuadamente la diversidad lingüística y dialectal del idioma español. Es decir, ciertos dialectos son excluidos o invisibilizados causando que se promueva una "pseudolengua" o un idioma neutro artificial, lejos de los usos reales de las comunidades hablantes. Específicamente se detalla que son incapaces de: 1) Diferenciar entre las zonas dialectales convencionalmente aceptadas (español caribeño, español mexicano y centroamericano, español andino, español rioplatense y español chileno). 2) Abarcar contextos y registros comunicativos formales e informales vinculados a diferentes variedades (variación sociolectal). 3) Adaptarse a las distintas tipologías textuales y contextos comunicativos (variación estilística y pragmática).

Ante este problema el articulo revisó diferentes iniciativas nacionales (incluyendo la Hoja de ruta de Colombia), y se destaca que, si bien hay enfoques locales, existe una urgencia por lograr una acción coordinada con datos de cada variedad geográfica. Pero no cualquier corpus, se debe cumplir con que: 1) Sean suficientes datos procedentes de las principales variedades lingüísticas; 2) Sean datos abiertos y disponibles para la investigación y el uso comercial; y 3) Sean datos con calidad suficiente para el entrenamiento de modelos de IA.

Es decir, este artículo muestra que las variantes dialectales del español, como lo son los modismos colombianos, son excluidas de los LLM y generan SLD; y es necesario una base de conocimiento que represente de manera adecuada esta información, como es el fin de este estudio. Porque, incluso modelos como MarIA, un modelo pre entrenado específicamente para el idioma español por el Centro Nacional de Supercomputación de Barcelona (CNS), se basó en corpus constituido como: 50% de español de España, 12% de español de México, 7% de español de Estados Unidos, y solo un 5% para el español de Colombia, Argentina y Chile cada uno. Esto es aún más grave para modelos comerciales conocidos donde cuentan con una base en lengua inglesa que, en algunos casos, puede llegar al 90 % de su corpus documental y que se convierten al español mediante traducción automática.

El segundo concepto, luego de cubrir el estado del arte respecto a los datos usados para entrenamiento del idioma español, se aborda en "*Investigating Idiomaticity in Word Representations*" [10] (Anexo[1] 2) y busca entender si la representación sobre la cual corren los LLM captura o no los modismos. El estudio evalúa la capacidad de los modelos de representación (como Word2Vec o BERT) para capturar la idiomática de compuestos nominales en inglés y portugués; donde, la principal conclusión es que los modismos aún no están representados con precisión.

La evaluación se hizo sobre el conjunto de datos: Noun Compound Idiomaticity Minimal Pairs (NCIMP) Dataset, que contiene pares (modismos compuestos por dos palabras) diseñados para probar la sensibilidad de los modelos a las perturbaciones léxicas que cambian el significado idiomático. Además, se proponen dos métricas que funcionan independientes al modelo: Affinity, una medida comparativa de similitud entre el elemento deseado y un distractor; y Scaled Similarity (SimR), que re escala la similitud para distinguir las coincidencias significativas de los ejemplos usados al evaluar el modelo.

Los resultados mostraron que, los modelos priorizan las pistas léxicas proporcionadas por los pares de palabras que las definiciones figurativas de las mismas. Esto se evidencia dado que al hacer sustituciones aleatorias o sustituciones de sinónimos literales de los componentes se obtenían resultados de alta similitud. Lo que sugiere que la cantidad de palabras compartidas entre las dos frases domina sobre la comprensión semántica.

Es decir, el significado de la frase viene de la combinación de los significados individuales de las palabras que la componen, y no del entendimiento real del modismo. Incluso los modelos contextualizados (como BERT y Llama2) no mostraron una superioridad clara sobre los modelos estáticos en esta tarea, lo que sugiere que la información contextual no se incorpora adecuadamente para el modismo; y hay una limitada capacidad de los LLM para interpretarlos de manera correcta.

Por tanto, con base en los resultados del estudio, hay que tener dos puntos presentes al momento de definir el concepto de similitud en nuestro proyecto: I) No podemos comparar la representación vectorial del modismo contra sus sinónimos directamente, dado que como vimos, los modelos de representación no capturan el significado de manera adecuada. II) Comparar dos frases (el ejemplo del modismo en nuestro dataset y otra sustituyendo el mismo usando un LLM), no son un buen estimador para saber si hay un entendimiento del lenguaje figurativo.

Finalmente, se procede a revisar dos baselines y sus metodologías usadas, que cuantifican el entendimiento de los LLM sobre los modismos. El primer baseline parte del estudio "*Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection*" [12] (Anexo[1] 3) donde se evalúa el rendimiento de varios LLM, incluyendo tanto implementaciones locales como modelos de software-como-servicio (SaaS), en la tarea de detección de modismos. Esta evaluación se realiza sobre tres conjuntos de datos: SemEval 2022 Task 2a, MAGPIE y la porción de modismos de FLUTE; y responde a la pregunta: ¿la observación es un modismo? Con posibles respuesta: idiomático o literal.

La idea principal fue entonces determinar si los LLM pueden igualar el desempeño de los modelos encoder-only más pequeños que han sido afinados específicamente para

[2] Contiene tipo de modelo entrenado, datos utilizados, y metodología de evaluación.

estas tareas. Los modelos evaluados incluyen los modelos SaaS de OpenAI y Google, junto con modelos locales como Llama2, Phi-2, Mistral-7B, y otros.

Los resultados muestran que, a pesar de que los LLM a mayor escala ofrecen un rendimiento competitivo en la detección de modismos, no logran igualar los resultados alcanzados por los modelos encoder-only afinados y específicos para la tarea. Por ejemplo, en la tarea SemEval 2022 Task 2a en la configuración zero-shot, el mejor rendimiento fue de 0.890 F1, muy superior a los resultados obtenidos por los modelos GPT-4.

Se debe resaltar que, el rendimiento en la detección de modismos mostró una tendencia consistente de escalamiento con el número de parámetros del modelo. Esta tendencia fue evidente en las variantes de Llama2, donde 13B superó a 7B, y en los modelos Flan-T5. Se resalta además el uso de experimentos de prompt engineering con el fin de observar si existían mejoras en los resultados. Se encontró que la personificación de expertos, como "Experto en uso del lenguaje", no mejoró el rendimiento de GPT-3.5-turbo en el subconjunto de SemEval en inglés. Pero, el análisis multilingüe de SemEval, al especificar de manera explícita del idioma y la traducción del prompt mejoraron el rendimiento para el gallego, un idioma presumiblemente menos representado en los datos de entrenamiento de los modelos.

Lo cual da a nuestro estudio un primer marco metodológico a replicar: dado que en nuestro dataset contamos con ejemplos que hacen uso del significado figurativo del modismo, podemos preguntar a los diferentes LLM la misma pregunta de este estudio con el fin de calcular las mismas métricas. Así mismo se puede pensar, dependiendo de los resultados, en especificar de manera explícita la variante dialéctica, dado que funciono con el gallego.

Inclusive, este primer baseline da como tema de discusión / trabajo futuro, el entrenamiento especializado de un encoder-only y la comparación con estás métricas preliminares para confirmar si los modelos actuales aún se desempeñan peor que un modelo especializado. De igual manera, el segundo baseline "*It's not Rocket Science: Interpreting Figurative Language in Narratives*" [11] (Anexo[1] 4) abre las puertas a otro debate interesante.

Este artículo estudia la comprensión de modismo y símiles en el procesamiento del lenguaje natural, resaltando la limitación fundamental de los LLM basados en Transformers, como GPT, al intentar comprender el lenguaje figurado; pues, si bien son excelentes para procesar palabras dentro de su contexto, su método de composición no es adecuado para capturar el significado figurado o no literal que surge cuando el significado de la frase es diferente al de las palabras que la componen.

El artículo usa una base de datos de modismos construido de manera manual usando el Toronto Book corpus, que es una colección de 11,038 libros electrónicos independientes; junto con 117 trabajadores de Amazon Mechanical Turk que escribieron continuaciones de los contextos (los cuales tenían modismos presentes). La base de datos final de modismos contiene 5,101 tuplas <narrativa, continuación>, sin incluir ninguna expresión que se utilice de la forma literal. Es decir, por ejemplo: "running a mile" puede significar I) evitar algo de cualquier manera posible, o II) correr una milla. En la base de datos final se conserva solo el primer significado.

Este, lo dividieron en los siguientes sets: 3,204 para entrenamiento, 355 para validación y 1542 para prueba. Se dividieron los ejemplos de tal forma que no existieran modismos repetidos entre los sets de entrenamiento y de prueba, logrando probar la habilidad de los modelos para generalizar a modismos no vistos.

El estudio evaluó la capacidad de diversos LLM para interpretar el lenguaje figurado en narrativas. Se definieron dos tareas principales: discriminativa y generativa. La discriminativa consistía en, dada la expresión figurativa, seleccionar la continuación plausible entre dos opciones posibles y la generativa en generar una continuación plausible que sea coherente con la narrativa y se ajuste al significado de la expresión figurativa.

Los resultados mostraron que los LLM pre entrenados, incluidos GPT-3, tuvieron un rendimiento sustancialmente peor que los humanos, con brechas de precisión de hasta 14.6 puntos en la tarea discriminativa y hasta 28 puntos en la evaluación de plausibilidad generativa. No obstante, el estudio demostró que la introducción de modelos mejoradores de conocimiento, inspirados en estrategias humanas (inferir del contexto y usar el significado literal de los constituyentes), mejoran consistentemente el desempeño.

Es decir, si bien la metodología no es replicable para nuestro proyecto dada la falta de continuaciones para nuestros ejemplos, la afirmación de mejoras en el desempeño por medio de la introducción de conocimiento a partir de modelos como COMET, ponen sobre la mesa una discusión clara: si ningún modelo es capaz de reconocer los modismos colombianos, usando COMET para generar inferencia sobre los ejemplos, es posible mejorar el entendimiento de un dialecto no dominante desde los datos de entrenamiento, o es una metodología únicamente replicable con el inglés.

Fuera de los trabajos futuros y discusiones que se han abierto con la revisión del estado del arte respecto al entendimiento de los modismos por parte del LLM, se pudo observar de manera clara que la problemática no es solo cuestión de la arquitectura del modelo, pues viene desde los datos de entrenamiento y la representación de estos. Por lo cual, no solo es relevante la creación de una base de conocimiento adecuada de modismos como se busca con este proyecto, sino la definición de una metodología clara y replicable.

IV. METODOLOGÍA PROPUESTA

Vamos a utilizar los siguientes modelos con base en los baseline discutidos en la revisión de literatura:

Modelos GPT y Gemini: GPT-5, GPT-4, Gemini 2.5 Pro

Modelos Llama y Phi: Phi-4 (14.7B), Llama 4 Scout (109B), Llama 4 Maverick (400B)

Modelos Flan-T5: Flan-T5-XL, Flan-T5-XXL (11.3B)

La metodología que vamos a usar para evaluar los modelos se divide en tres, partiendo de los datos disponibles en nuestra base de conocimiento:

1. **Entendimiento de la definición:** tomaremos la palabra y le pediremos a cada uno de los LLM su definición más probable para la geografía colombiana. Tanto la definición original, como la obtenida del LLM, pasan por un proceso de embedding (concatenando las representaciones de cada palabra de la definición). Con estos vectores, realizaremos similitud por coseno, y evaluaremos si el modelo posee conocimiento sobre la variante en español con base a si da valores cercanos a uno o no. Vale la pena resaltar que no es necesario usar modificaciones de la métrica de similitud por coseno como paso en la revisión de literatura, dado que, al usar la definición de la palabra, la sobreposición léxica si es informativa para nosotros.

2. **Interpretación del ejemplo:** le vamos a pedir a cada LLM que responda si el ejemplo hace uso de modismo o no (no modismo). Con base a sus respuestas, podemos calcular que porcentaje de los ejemplos los clasifico de manera correcta. Partimos del hecho de que todos los ejemplos de nuestra base de conocimiento hacen uso del sentido figurado en el ejemplo, es decir, no es necesario un etiqueta humano sobre la tarea de clasificación dado que todos caen en la categoría “modismo”. Con el fin de evitar que aprenda la clase y responda siempre lo mismo (dada la falta de ejemplos negativos), se resalta que no se realizará ningún tipo de fine-tuning sobre ninguno de los modelos y cada consulta se hará de manera independiente a las anteriores.

3. **Consistencia del modelo:** le vamos a dar el ejemplo que contiene el modismo y le pediremos que diga cuál es la palabra más probable por la que remplazaría el modismo. Luego, dada esa nueva palabra, le pedimos al LLM la definición, y nuevamente comparamos la representación vectorial de la definición dada por el LLM y la definición original del modismos. De esta manera evaluamos si el modelo puede desambiguar y traducir figurativamente una expresión informal a su equivalente literal o estándar.

Vale la pena resaltar para los puntos 1 y 3 que, dado que usaremos la representación de las definiciones, y no la representación del modismo / sinónimo, la similitud por coseno es potencialmente una buena métrica de comparación. Además, sí las definiciones siguen siempre un uso literario de las palabras, creemos adecuado usar en primera instancia una representación estática como Word2vec, y dependiendo de los resultados una contextual como BERT.

Los resultados de los puntos 1 y 3 se evaluarán bajo el umbral de 0.5/1.0. Es decir, si la media de un modelo está por encima del umbral se considera que entiende modismos; de lo contrario no. Con el fin de ser ilustrativos, se graficarán los resultados en un diagrama de cajas y bigotes, con los resultados de cada modelos de manera independiente.

Para el punto 2, únicamente se reportarán las métricas de precisión, accuracy, y F1 en formato tabla para cada modelo; sujeto a revisión dada la falta de ejemplos negativos. Finalmente, todos los modelos se trabajarán bajo su configuración de Zero-shot.

V. BIBLIOGRAFÍA

[1] W. Liang, Y. Zhang, M. Codreanu, J. Wang, H. Cao y J. Zou, «The widespread adoption of large language model-assisted writing across society,» *Patterns,* p. 101366, 2025.

[2] N. Kobie, «What idioms teach us about AI and its evolution,» *PC Pro,* p. 126–128, August 2025.

[3] M. Mayor-Rocher, C. Pozo, N. Melero, G. Martínez, M. Grandury y P. Reviriego, «It's the same but not the same: Do LLMs distinguish Spanish varieties?,» arXiv, Online, 2025.

[4] I. C. y. Cuervo, «Diccionario de colombianismos,» Lexicc / Instituto Caro y Cuervo, 2024. [En línea]. Available: https://lexicc.caroycuervo.gov.co/diccionario/diccionario-de-colombianismos/.

[5] A. C. d. l. Lengua, «Base de datos de la Academia Colombiana de la Lengua,» s.f.. [En línea]. Available: https://www.academiacolombianadelalengua.co/bdc/.

[6] S. Rivas, «Presentación del proyecto: Diccionario de colombianismos,» Desparchado, 26 abril 2018. [En línea]. Available:

https://desparchado.co/events/diccionario-de-colombianismos/.

[7] KienyKe, «Colombianismos: un diccionario del español hablado,» KienyKe, 15 abril 2018. [En línea]. Available: https://www.kienyke.com/historias/colombianismos-un-diccionario-del-espanol-hablado.

[8] I. C. y. Cuervo, «Sobre el proyecto – Diccionario de colombianismos,» Diccionario de colombianismos, s.f.. [En línea]. Available: https://colombianismos.caroycuervo.gov.co/challenge.

[9] I. C. y. Cuervo, «Historia del Instituto Caro y Cuervo,» Instituto Caro y Cuervo, 28 marzo 2024. [En línea]. Available: https://www.caroycuervo.gov.co/instituto/historia/.

[10] W. He, T. K. Vieira, M. Garcia, C. Scarton, M. Idiart y A. Villavicencio, «Investigating Idiomaticity in Word Representations,» *Computational Linguistics,* vol. 51, nº 2, p. 505–555, 2025.

[11] T. Chakrabarty, Y. Choi y V. Shwartz, «It's not Rocket Science: Interpreting Figurative Language in Narratives,» *Transactions of the Association for Computational Linguistics ,* vol. 10, p. 589–606, 2022.

[12] D. Phelps, T. Pickard, M. Mi, E. Gow-Smith y A. Villavicencio, «Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection,» arXiv, Online, 2024.

[13] J. Muñoz-Basols, M. d. M. Palomares Marín y F. Moreno Hernández, «El Sesgo Lingüístico Digital (SLD) en la inteligencia artificial: implicaciones para los modelos de lenguaje masivos en español,» *Lengua y Sociedad. Revista de Lingüística Teórica y Aplicada,* vol. 23, nº 2, p. 623–647, 2024.

VI. ANEXOS

ANEXO 1 - El Sesgo Lingüístico Digital (SLD) en la inteligencia artificial: implicaciones para los modelos de lenguaje masivos en español. [13]

Los modelos entrenados incluyen diversas arquitecturas lingüísticas: I) MarIA, un LLM para español basado en GPT-2 y RoBERTa. II) BETO y BERTIN, modelos monolingües impulsados de manera independiente. III) Modelos multilingües como ILENIA y AINA, que incluyen versiones como Ăguila 7B y FLOR 6.3B (Catalán, Español, Inglés), además de modelos específicos para Gallego (Carballo) y Euskera (LATXA).

En cuanto a los datos utilizados, el corpus BNE incluye un 50% de español de España, 12% de México y solo 5% de Colombia, Argentina y Chile cada uno. Los modelos multilingües emplean datasets especializados como C4_es, BOE, CSIC, Eurlex y Wikipedia, entre otros, a menudo con porcentajes significativos de inglés.

La metodología de evaluación se basó en dos componentes principales. Primero, una revisión documental que abarcó el análisis de documentos gubernamentales, estrategias nacionales, notas de prensa de proyectos oficiales y artículos académicos. Segundo, un estudio de modelos abiertos mediante el uso de la plataforma HuggingFace y sus respectivas citaciones académicas.

Entre las conclusiones principales, se destaca la necesidad de una acción coordinada para reflejar y preservar la diversidad lingüística del español, a través de datos de calidad representativos utilizados para futuros entrenamientos de los modelos. Asimismo, se resalta la importancia de mitigar el SLD, ya que la falta de representatividad promueve una "pseudolengua" o "dialectos digitales" que no reflejan de manera fiel a los hablantes reales.

ANEXO 2 - Investigating Idiomaticity in Word Representations [10]

El estudio evaluó modelos de representación de palabras pre entrenados con arquitecturas estáticas y/o contextualizadas. Entre los modelos estáticos se encuentran Word2Vec y GloVe, mientras que contextualizados se consideraron ELMo, BERT (en sus versiones monolingüe y multilingüe —mBERT—), DistilBERT multilingüe (mDistilB), Sentence-BERT multilingüe (mSBERT) y Llama2.

Los datos utilizados provienen del Noun Compound Idiomaticity Minimal Pairs (NCIMP) Dataset, que contiene 32,200 frases con compuestos nominales (NCs) clasificados por nivel de idiomaticidad: idiomáticos, parcialmente composicionales y composicionales. El conjunto de datos abarca dos lenguajes, inglés (280 NCs, 19,600 frases) y portugués (180 NCs, 12,600 frases), e incluye contextos naturalísticos (Nat, informativos) y semánticamente neutrales (Neut, no informativos). Además, las anotaciones incluyen juicios humanos de idiomaticidad en una escala Likert de 0 (totalmente idiomático) a 5 (totalmente composicional), junto con paráfrasis o sinónimos de referencia (gold standard synonyms) para los NCs.

La metodología de evaluación se basa en una estrategia de sondeo que crea pares mediante la sustitución del

compuesto nominal objetivo por diferentes palabras. Las variantes son: PSyn, que sustituye por el sinónimo de referencia; PWordsSyn, que sustituye cada palabra por un sinónimo literal; PComp, que reemplaza solo una de las palabras; y PRand, que sustituye por una expresión aleatoria controlada por frecuencia.

Las métricas clave empleadas para evaluar las representaciones vectoriales fueron: Similitud (Sim), que mide la similitud por coseno entre la frase original y la modificada; Affinity (Aff), que evalúa si la representación objetivo es más similar a un ejemplo relacionado (como un sinónimo) que a uno no relacionado (como una sustitución aleatoria); Scaled Similarity (SimR), que reescala la similitud tomando como límite inferior las sustituciones aleatorias (equivalente a una normalización max-min en el espacio no uniforme del modelo); y Correlación (ρ), que utiliza la correlación de Spearman para medir la relación entre las métricas de similitud y los juicios humanos de composicionalidad. Las representaciones vectoriales se obtuvieron mediante la concatenación de los embeddings individuales.

Entre las conclusiones principales, se encontró que los modelos, en general, no lograron capturar con precisión la idiomaticidad, ya que no se observó el patrón esperado de similitudes entre el compuesto nominal y sus sinónimos. Los resultados de Affinity y Scaled Similarity mostraron que, para los compuestos más idiomáticos, los modelos son más sensibles a las sustituciones que preservan la forma literal (PWordsSyn) que a aquellas que preservan el significado idiomático (PSyn). Esto indica que los modelos se basan en los significados de las palabras individuales en lugar de incorporar un significado figurativo. Finalmente, los modelos contextualizados no demostraron una ventaja significativa sobre los modelos estáticos en la representación de la modismos.

ANEXO 3 - Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection [12]

Los LLMs como GPT-3.5-turbo, GPT-4, GPT-4-turbo y Gemini 1.0 Pro, fueron entrenados utilizando tres datasets de expresiones idiomáticas en contexto: SemEval 2022 Task 2a, FLUTE y MAGPIE. La evaluación del rendimiento de cada modelo se realizó mediante la puntuación macro-average F1, utilizando zero-shot prompting por defecto. Resaltamos, la exploración de ingeniería de prompts con la impersonación de expertos y la especificación de formato, así como el few-shot prompting en configuraciones one-shot y few-shot. Los resultados mostraron que los LLM a mayor escala (en número de parámetros), como GPT-4, alcanzan un desempeño competitivo en la detección de modismos; aunque, no igualan el rendimiento de modelos encoder-only mucho más pequeños a los que se les ha hecho fine-tuned para tareas específicas sobre SemEval y MAGPIE.

Los Modelos Abiertos Locales, como Llama2 (7B, 13B), Phi-2 y CapybaraHermes-2.5-Mistral-7B, se entrenaron con los mismos datasets pero de manera local utilizando variantes cuantizadas (Q5_K_S) con el fin de minimizar el impacto en el rendimiento y hacerlos compatibles con hardware de consumo. La evaluación se realizó mediante zero-shot prompting. Los resultados, nuevamente indican que el rendimiento de los LLM generativos y asistidos por prompting escala de manera consistente con el número de parámetros para estos datasets, como se evidencia en Llama2 y Flan-T5, lo que sugiere un potencial de mejora en modelos de mayor tamaño.

Finalmente, los Modelos Multilingües, representados por Flan-T5 en sus versiones Small, Base, Large, XL y XXL, fueron evaluados con los datasets SemEval 2022 Task 2a (EN, PT-BR, GL). La metodología consistió en analizar el impacto del tamaño del modelo en el rendimiento, evaluar el desempeño multilingüe y realizar pruebas extensivas de one-shot y few-shot prompting. Los resultados muestran que, para idiomas menos representados en los datos de entrenamiento (como el gallego), especificar el idioma objetivo y/o traducir los prompts puede mejorar el rendimiento.

ANEXO 4 - It's not Rocket Science: Interpreting Figurative Language in Narratives [11]

Los LLM incluyeron configuraciones zero-shot (GPT-2 XL, GPT-3, UnifiedQA), few-shot (GPT-3, PET) y supervisadas (RoBERTa-large, T5-large, BART-large, GPT-2 XL). Estos modelos se entrenaron con narrativas cortas del Toronto Book Corpus que contenían modismos o símiles. Además, se utilizaron dos datasets: uno de modismos (5,101 tuplas) y otro de símiles (4,996 tuplas), cada uno con continuaciones verosímiles e inverosímiles obtenidas mediante crowdsourcing. Los datos se dividieron cuidadosamente para asegurar que no existiera superposición de expresiones figuradas entre los conjuntos de entrenamiento y prueba.

La metodología de evaluación consistió en aplicar el formato Story Cloze Test para las siguientes dos tareas: (1) discriminativa, en la que el modelo debía elegir la continuación más plausible, y (2) generativa, en la que debía generar una continuación plausible. El rendimiento humano alcanzó un 92.0% en modismos y un 95.0% en símiles. Las métricas utilizadas fueron precisión (accuracy) para la tarea discriminativa, y ROUGE-L junto con BERT-Score para la generativa. Además, se realizaron evaluaciones humanas mediante juicios de plausibilidad absolutos y comparativos. Los resultados mostraron que los modelos de lenguaje base tuvieron un rendimiento sustancialmente inferior al humano, con una brecha de hasta 14.6 puntos en precisión discriminativa y 28 puntos en plausibilidad generativa. Los modelos zero-shot y few-shot (incluido GPT-3) mostraron

bajo rendimiento, evidenciando que no son capaces de entender modismos.

Por otro lado, los Modelos Mejorados con Conocimiento incluyeron dos variantes: el Modelo Literal y el Modelo de Contexto. Ambos utilizaron los mismos datasets base, pero enriquecidos con conocimiento proveniente de COMET-ConceptNET (para representar el significado literal) y ParaCOMET-ATOMIC (para inferencias contextuales y de eventos). Las inferencias de conocimiento se integraron mediante concatenación del input en las tareas discriminativas y mediante ensamblaje de logits en las tareas generativas. Manteniendo las mismas métricas y la metodología descrita, los resultados mostraron que estos modelos superaron consistentemente a los modelos base en ambas tareas, reduciendo la brecha con el rendimiento humano en el caso de los modismos.