Lengua y Sociedad. Revista de Lingüística Teórica y Aplicada,
Vol. 23, n.º 2, julio-diciembre 2024, pp. 623-647,
ISSN-L 1729-9721; eISSN: 2413-2659
https://doi.org/10.15381/lengsoc.v23i2.28665

El Sesgo Lingüístico Digital (SLD) en la inteligencia artificial: implicaciones para los modelos de lenguaje masivos en español

The Digital Linguistic Bias (DLB) in Artificial Intelligence: Implications for Large Language Models in Spanish

O Viés Linguístico Digital (VLD) na Inteligência Artificial: implicações para grandes modelos de linguagem em espanhol

Javier Muñoz-Basols
Universidad de Sevilla, España / University of Oxford, Reino Unido
javier.munoz-basols@mod-langs.ox.ac.uk
https://orcid.org/0000-0003-3856-3637

María del Mar Palomares Marín
University of Limerick, Irlanda
maria.palomares@ul.ie
https://orcid.org/0000-0002-8474-3375

Francisco Moreno Fernández
Observatorio Global del Español, Instituto Cervantes, España - Universität Heidelberg, Alemania
francisco.moreno@uni-heidelberg.de
https://orcid.org/0000-0002-3136-4443

Resumen

La llegada de la inteligencia artificial generativa a nivel de usuario, especialmente a partir de los Modelos de Lenguaje Masivos (MLM), nos obliga a reflexionar sobre la proliferación de sesgos en la construcción, desarrollo, uso y representatividad de estos modelos basados en datos lingüísticos. En este artículo, se revisan las iniciativas desarrolladas para el español en el campo de la inteligencia artificial (IA), tanto desde la América hispanohablante como desde España, de modo que se presta especial atención a los recursos lingüísticos y a los MLM. Se examina la composición de los principales MLM actuales del español y se comparan con otros MLM de lenguas peninsulares (catalán, euskera, gallego y valenciano). Asimismo, se introduce el término Sesgo Lingüístico Digital (SLD) para identificar la hibridez lingüística que la IA genera tanto a nivel interlingüístico (p. ej., en relación con la base del inglés utilizada para entrenar estos modelos) como intralingüístico (en relación con las distintas variedades de la lengua). Finalmente, se sugiere que un usuario con conciencia digital podrá contribuir a mitigar los efectos del SLD. En conclusión, se enfatiza la necesidad de una acción coordinada por parte de los agentes institucionales para preservar la diversidad del patrimonio lingüístico hispanohablante en el desarrollo de los MLM.

Palabras clave: Inteligencia Artificial (IA); Modelos de Lenguaje Masivos (MLM); Sesgo Lingüístico Digital (SLD); diversidad de la lengua; variación dialectal; español.

Abstract

The advent of generative artificial intelligence at the user level, particularly through the development of Large Language Models (LLMs), prompts us to reflect on the proliferation of biases in the construction, development, use, and representation of these models based on linguistic data. This article first reviews the initiatives developed for Spanish in the field of AI from Latin America and Spain, with special attention to linguistic resources and LLMs. The composition of the current major LLMs for Spanish is examined and compared with other LLMs for peninsular languages (Catalan, Basque, Galician, and Valencian). Subsequently, the term Digital Linguistic Bias (DLB)is introduced to identify the linguistic hybridity generated by AI both at an interlinguistic level (e.g., in relation to the English base used to train these models) and an intralinguistic level (in relation to the different varieties of the language). Finally, it is suggested that a digitally aware user can intervene mitigating the effects of the DLB. In conclusion, the need for coordinated action by institutional agents to preserve the diversity of the Spanish-speaking linguistic heritage in the development of LLMs is emphasized.

Keywords: Artificial Intelligence (AI); Large Language Models (LLMs); Digital Linguistic Bias (DLB); language diversity; dialectal variation; Spanish..

Resumo

O advento da inteligência artificial generativa no nível do usuário, especialmente por meio do desenvolvimento de Grandes Modelos de Linguagem (GML), nos leva a refletir sobre a proliferação de vieses na construção, no desenvolvimento, no uso e na representatividade desses modelos baseados em dados linguísticos. Este artigo analisa, em primeiro lugar, as iniciativas desenvolvidas para o espanhol no campo da IA, tanto na América de língua espanhola quanto na Espanha, dando atenção especial aos recursos linguísticos e aos GML. A composição dos principais GML atuais do espanhol é examinada e comparada com outros GML de idiomas peninsulares (catalão, basco, galego e valenciano). Além disso, o termo Viés Linguístico Digital (VLD)é apresentado para identificar a hibridez linguística gerada pela IA, tanto em nível interlinguístico (por exemplo, em relação à base em inglês utilizada para treinar esses modelos) quanto em nível intralinguístico (em relação às diferentes variedades da língua). Por fim, sugere-se que um usuário digitalmente consciente poderá contribuir para atenuar os efeitos do VLD. Para concluir, enfatiza-se a necessidade de uma ação coordenada dos agentes institucionais para preservar a diversidade do patrimônio linguístico de língua espanhola no desenvolvimento de GML.

Palavras chave: Inteligencia Artificial (AI); Grandes Modelos de Linguagem (GML); Viés Linguístico Digital (DLB); diversidade linguística; variação dialetal; espanhol.

Recibido: 01/05/2024 Aceptado: 27/07/2024 Publicado: 30/12/2024

1. Introducción

La llegada de la inteligencia artificial (IA) generativa ha traído consigo numerosas implicaciones sociales, especialmente, desde que la accesibilidad a nivel de usuario ha permitido la interacción con herramientas potenciadas por esta tecnología. Una de las consecuencias más inmediatas a nivel institucional ha sido la creación de estrategias y hojas de ruta nacionales para la integración de la IA, no solo para gestionar su implementación en los próximos años, sino también para analizar cómo optimizar el uso de esta nueva tecnología. En este contexto, una serie de países de habla hispana han publicado documentos oficiales desde los que se aborda este nuevo entorno tecnológico y se explora cómo la IA puede contribuir al desarrollo de diversos sectores de la sociedad.

El *Plan nacional de inteligencia artificial* de Argentina (Presidencia de la Nación, 2020) busca promover un crecimiento sostenible y mejorar la igualdad de oportunidades en el país. En Chile, la *Política nacional de inteligencia artificial* (Arancibia *et al*., 2021; MinCiencia, 2024) destaca por la infraestructura para datos, la investigación sobre la materia y la conectividad. Por su parte, Colombia ha presentado la *Hoja de ruta para el desarrollo y aplicación de la inteligencia artificial* (Ministerio de Ciencia, Tecnología e Innovación, 2024), que incluye el programa 'ColombIA Inteligente', orientado a la ejecución de proyectos de investigación aplicada e innovación basados en IA y tecnologías aeroespaciales.

En México, la *Propuesta de agenda nacional de la inteligencia artificial* (Lagunes *et al.*, 2024), elaborada por la Alianza Nacional de Inteligencia Artificial (ANIA), es fruto de un grupo de trabajo con iniciativa privada, académica, pública y organismos internacionales. Perú ha puesto en marcha su *Estrategia nacional de inteligencia artificial* (Presidencia del Consejo de Ministros, 2021) enfocada en el ámbito económico y social del país. Uruguay, con su *Estrategia de inteligencia artificial para el gobierno digital* (Agesic, 2019), busca priorizar la transparencia en el uso de algoritmos y la protección de datos. Por último, España, como parte de su *Estrategia de inteligencia artificial* (Gobierno de España, 2024), está trabajando en un modelo fundacional para el español.

Pese a la importancia y complementariedad de estas propuestas, iniciativas y estrategias nacionales, en estos documentos destaca predominantemente un "enfoque local" que no considera de manera integral y estratégica —al menos desde el punto de vista lingüístico— las implicaciones necesarias en entornos de IA para una lengua sumamente diversa como el español. Así, la creación de Modelos de Lenguaje Masivos (MLM) (*Large Language Models*, LLMs) en los países hispanohablantes debería aspirar a una "estrategia global" que reconozca el notable grado de cohesión y homogeneidad de una lengua compartida por más de 500 millones de personas, pero con una configuración pluricéntrica y sin una única variedad estándar, características que deben ser reflejadas en todos los ámbitos de la comunicación, incluyendo el desarrollo de la IA.

Los MLM se alimentan de datos lingüísticos, lo que subraya la importancia de analizar las implicaciones que se desprenden de su creación desde la fase de desarrollo de corpus y la incorporación de modelos textuales, de voz y de traducción, pasando por la fase de transformación y distribución, como el despliegue de modelos especializados (*Small Language Models*, SLM) o el diseño de aplicaciones, hasta la fase de uso y/o consumo de estos modelos por parte de empresas, ciudadanos o del sector público (Gobierno de España, 2024). Aunque está previsto el lanzamiento de ALIA, un gran modelo fundacional abierto de IA generativa para el español (Gobierno de España, 2024), que, entre otros aspectos, no dependa de datos en inglés para su entrenamiento, aún persisten numerosas incógnitas sobre cómo se desarrollará este modelo de manera que represente adecuadamente la diversidad lingüística del mundo hispanohablante. Será necesario "garantizar una presencia digital del español que no sea monolítica sino rica, diversa y dinámica" (García Montero, 2022, p. 144).

2. La IA y la lengua española

El *Índice latinoamericano de inteligencia artificial (ILIA)* (VV. AA., 2023), creado por el Centro de Estudios Nacional de Inteligencia Artificial del Gobierno de Chile para medir la infraestructura e impacto de la IA en Argentina, Bolivia, Brasil, Chile, Colombia, Costa Rica, Ecuador, México, Panamá, Paraguay, Perú y Uruguay, y elaborado igualmente con datos de la OCDE, la Cepal y la Unesco, permite identificar algunas de las necesidades inmediatas para la consolidación de la IA en estos contextos geográficos. El índice abarca diversas dimensiones del impacto de la IA en la sociedad, tales como los factores habilitantes; la investigación, el desarrollo y la adopción de la IA; la gobernanza; la percepción de la IA y el futuro de la IA. Es decir, proporciona una visión actual de la infraestructura tecnológica existente, identifica los contextos más preparados para esta nueva era tecnológica e investiga las percepciones de los ciudadanos y plantea perspectivas de futuro.

Este documento destaca, asimismo, la capacidad de los contextos latinoamericanos para aprovechar en mayor o menor medida la IA generativa y, por otro lado, explica las diferentes estrategias nacionales. En definitiva, la IA generativa es una tecnología disruptiva (Dafoe, 2018) que está transformando el panorama social en el espacio hispanohablante a todos los niveles, aunque con diferentes grados de adecuación a esta nueva realidad. Según Peláez Agudo (2023, p. 18), "la penetración relativa de habilidades tecnológicas y disruptivas, asociadas a la IA, es menor en América Latina (2,16 %) que en el resto del mundo (3,59 %), lo que aún limita este desarrollo en comparación con otras regiones del mundo".

Junto al aspecto tecnológico, es necesario tener en cuenta el modo en que interactúan los usuarios con esta nueva tecnología. El índice *ILIA* (VV. AA., 2023, p. 152) examina la percepción de la IA por parte de la población hispanohablante americana y destaca sus principales nudos críticos: la invasión de la privacidad, el daño socioambiental, la ética, la manipulación y los desórdenes informativos, la discriminación de género, el reemplazo y precarización laboral, la transparencia algorítmica y la responsabilidad y la discriminación ético-racial. Por ejemplo, sabemos que la IA perpetúa la discriminación de género en diversas tareas, contribuyendo a la reproducción de estereotipos. Esto se observa en el procesamiento del lenguaje (como en la traducción o las sugerencias de búsqueda) y en la clasificación de perfiles laborales de mujeres que solicitan trabajos técnicos, donde los algoritmos tienden a degradar los currículos de las solicitantes debido a su género (VV. AA., 2023, p. 175). En resumen, este comportamiento de la IA revela uno de los muchos sesgos que es necesario identificar, investigar y mitigar.

La lengua no está libre de los sesgos inherentes a la IA, considerando especialmente que se trata de la principal materia prima que alimenta los MLM. Estos modelos, también conocidos como *grandes modelos de lenguaje*, *modelos de lenguaje de gran tamaño* o *a gran escala*, se nutren de una combinación de corpus textuales existentes y de datos recopilados de internet, como páginas web, libros, artículos, materiales académicos y otros contenidos textuales digitales en redes sociales. Además, incluyen textos de diversos dominios, como documentos jurídicos, informes financieros o publicaciones médicas (Amaratunga, 2023; Liu *et al*., 2024). A pesar de que en la IA generativa "la diversidad cultural y lingüística de la región puede ser una fuente de ventajas competitivas al permitir el desarrollo de soluciones adaptadas a sus peculiaridades" (Peláez Agudo, 2023, p. 18), los actuales MLM no logran reflejar adecuadamente esta diversidad lingüística y dialectal, sino que existe "una criba dialectal y sociolectal […] que incluye y visibiliza ciertos dialectos mientras que excluye e invisibiliza otros" (Company Company, 2019, p. 109). Estos modelos no contienen porcentajes suficientemente representativos de las particularidades e idiosincrasias del idioma; es más, 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 (Portal Administración Electrónica, 2024).

Si el trazado de isoglosas o líneas que dividen zonas geográficas con características dialectales específicas ya es difícil de realizar consistentemente dentro de un mismo territorio desde una metodología dialectológica, esta tarea resulta menos fiable si se deja en manos de las máquinas en las condiciones actuales. Así, la irrupción de la IA generativa en Latinoamérica, a nivel lingüístico, revela su incapacidad para 1) diferenciar entre las zonas dialectales convencionalmente aceptadas (español caribeño, español mexicano y centroamericano, español andino, español rioplatense y español chileno, cada una de las cuales agrupa a varios países y múltiples zonas dentro de un

mismo país [variación geolectal]) (Moreno Fernández, 2000); 2) abarcar contextos y registros comunicativos formales e informales vinculados a diferentes variedades (variación sociolectal); y 3) adaptarse a las distintas tipologías textuales y contextos comunicativos (variación estilística y pragmática). Estos tipos de variación constituyen tradiciones lingüísticas y discursivas consolidadas, con características y rasgos que reflejan la manera de comunicarse en sociedad de una comunidad concreta de hablantes.

Considerando lo anterior, la IA deja de ser un tema exclusivo de unos pocos especialistas, como indican Nguyen y Hekman (2022), para convertirse en un "hecho social total" (Marres, 2017, citado en VV. AA., 2023, p. 152) que afecta a la población de un país y, por ende, a los hablantes de la lengua. A pesar de que aún estamos en una fase de transición hacia esta nueva tecnología a nivel de usuario, se observa una falta de transparencia en los sistemas algorítmicos y en las aplicaciones impulsadas por IA (VV. AA., 2023). Por ello, es especialmente importante cuestionar sus implicaciones sociales para considerar qué aspectos debemos tener en cuenta con respecto al español, una lengua muy diversa y heterogénea, con una sólida implantación pluricéntrica tanto social como institucional, reforzada por una política lingüística panhispánica común que sirve de referencia entre las comunidades hispanohablantes (Real Academia Española [RAE] y Asociación de Academias de la Lengua Española [ASALE], 2004).

Para llevar a cabo esta investigación, se adoptó una metodología descriptiva y analítico-crítica que combina dos tipos principales de fuentes de información para la recopilación de datos. En primer lugar, la revisión de documentos gubernamentales, notas de prensa de proyectos oficiales, artículos académicos y congresos de comunidades especializadas en lingüística computacional. Se recopiló y analizó toda esta información para obtener una visión integral de las políticas y estrategias de desarrollo de países hispanohablantes como Argentina, Chile, Colombia, España, México, Perú y Uruguay.

Y, en segundo lugar, se tomó como referencia la plataforma HuggingFace (https://huggingface.co/), el mayor repositorio de modelos abiertos, se estudiaron MLM utilizando la información disponible. Esta plataforma proporcionó fichas detalladas de cada modelo, que incluían enlaces a artículos publicados fundamentalmente en el servicio de distribución de artículos académicos de STEM, arXiv, lo que permitió una comprensión profunda de las características y aplicaciones de cada modelo. La triangulación de todos estos datos permitió realizar una evaluación crítica y bien fundamentada de las tendencias actuales y líneas de actuación futuras de la IA en el ámbito hispánico.

3. Los proyectos de IA y los datos disponibles en español

España encabeza los esfuerzos para monitorizar la evolución del español en el creciente entorno de la IA generativa. Hoy en día, diversos proyectos estratégicos por parte de la Administración General y las Comunidades Autónomas, donde se hablan lenguas cooficiales, están ayudando a establecer las bases para estas nuevas tecnologías impulsadas por la IA. El Plan de Recuperación y Transformación y Resiliencia (PERTE, Nueva Economía de la Lengua) del Gobierno de España ha acogido varias iniciativas en el ámbito de la IA, tales como el Plan LEIA, el Plan ILENIA (que comprende los proyectos NEL-ENIA, NEL-GAITU, NEL-VIVES y NEL-NÓS), el Observatorio de Neologismos y Tecnicismos (Gobierno de España, 2022) o TeresIA (Gobierno de España, 2023). Existen otras muchas iniciativas con potencial para el desarrollo de la IA y específicamente

de los MLM en español (Instituto Cervantes, Ministerio de Economía y Transformación Digital, 2023), tales como el macrocorpus de español hablado PRESEEA (Proyecto para el Estudio Sociolingüístico del Español de España y América) (Moreno Fernández y Cestero Mancera, 2020; Moreno Fernández, 2022) o el CORPEN-FUNDACIÓN COMILLAS, para la creación de un corpus de español de los negocios, así como diferentes iniciativas para la creación de futuros corpus, como la liderada por la Secretaría de Estado de Digitalización e Inteligencia Artificial del Gobierno de España. Reconociendo su potencial, en este momento preferimos centrarnos en aquellos proyectos nacidos con el propio desarrollo de la IA (Moreno Sandoval, 2024).

Desde 2019, el plan LEIA (Lengua Española e Inteligencia Artificial) ha establecido acuerdos con importantes empresas tecnológicas como Amazon, Google, Facebook, Telefónica y Twitter (ahora "X"). Gracias a estos acuerdos, el *Diccionario de la lengua española (DLE)*, de la Real Academia Española (RAE), y la Asociación de Academias de la Lengua Española (ASALE), se han integrado como fuente de información en las búsquedas de Google y en el teclado de Google para dispositivos Android. Microsoft, por su parte, también ha obtenido acceso al *DLE* para todos sus productos, incluyendo el buscador Bing, Microsoft 365 y herramientas de revisión ortográfica y gramatical, que incluye tanto el español europeo como las variedades americanas con el fin de evitar sesgos lingüísticos en sus aplicaciones de IA (Real Academia Española, 2022a). Asimismo, Amazon, mediante su colección de servicios de computación en la nube, Amazon Web Services (AWS), ha creado una herramienta de análisis lingüístico del español llamada *Herramienta RAE-AWS*. Esta herramienta, potenciada por IA, puede examinar miles de documentos en internet y recopilar grandes volúmenes de datos, además de elaborar informes sobre temas lingüísticos, como el estudio de extranjerismos, el análisis de la riqueza léxica o la identificación y clasificación de errores lingüísticos (Real Academia Española, 2022b).

Por su parte, ILENIA (Impulso de las Lenguas en Inteligencia Artificial) es un ambicioso plan destinado a promover la lengua española y las lenguas cooficiales (catalán, euskera, gallego y valenciano) para crear recursos multilingües que ayuden a desarrollar tecnologías del lenguaje (asistentes de voz, agentes conversacionales, traducción automática, entre otros) (ILENIA, 2024).

En la Figura 1, se detalla una cronología de los principales proyectos impulsados desde España para la IA en español y, en la Tabla 1, se ofrece una síntesis de estos en función de su nomenclatura, descripción, objetivos y entidades colaboradoras. Estos proyectos abarcan aspectos lingüísticos esenciales para el avance de la IA en español, incluyendo la creación de corpus y bases de datos, así como el desarrollo de una terminología en español y de MLM.

Figura 1

Cronología de proyectos impulsados desde España para la IA en español

Tabla 1

Principales proyectos impulsados desde España para el desarrollo de la IA en español

Proyecto	Descripción	Objetivos	Entidades
LEIA (Lengua Española e Inteligencia Artificial [iniciado en 2019])	Uso correcto del español en el medio tecnológico.	- Uso correcto del español en las máquinas. - Herramientas de IA que promuevan el uso correcto de la lengua.	RAE y ASALE, Telefónica, Google, X, Microsoft, Amazon, Facebook.
MarIA (iniciado en 2021)	Primeros MLM para el español desarrollados con GPT-2.	- Crear MLM entrenados en español.	Centro Nacional de Supercomputación Barcelona (CNS) - Supercomputing Center (BSC), Biblioteca Nacional de España.
ILENIA (impulso de las Lenguas e Inteligencia Artificial). Dentro de ILENIA: - NEL-AINA (catalán) - NEL-VIVES (valenciano) - NEL-GAITU (euskera) - NEL-NÓS (gallego)	Impulsar en España la nueva economía digital basada en el lenguaje natural que promueva el español y las lenguas cooficiales (catalán, euskera, gallego y valenciano).	- Recursos multilingües de texto, voz y traducción automática en las lenguas cooficiales. - Primeros modelos multilingües en español de todas las lenguas cooficiales (Campusa, 2024).	Barcelona Supercomputing Center (BSC) (coordinación del proyecto). Centro de Inteligencia Digital Alicante (CENID), Universidad de Alicante. Centro de Tecnología de la Lengua (HiTZ), Universidad del País Vasco (UPV/EHU). CiTIUS, ILG, Universidad de Santiago de Compostela.
TeresIA (iniciado en 2023)	Terminología en español de alcance panhispánico.	- Portal de acceso a terminología en español. - Traducciones y redacciones de texto. - Terminologías de todo el ámbito panhispánico. - Corpus de literatura científica en español.	CSIC, Instituto Cervantes, CNS/BSC, Real Academia de Ingeniería, Asociación Española de Terminología, Grupo de Ontología de la Universidad Politécnica de Madrid, Dirección General de Traducción de la Comisión Europea.

En lo que a la disponibilidad de datos se refiere, esta resulta clave para avanzar en el progreso de la IA en español. La diversidad de las lenguas de los territorios hispanoamericanos, en sus manifestaciones digitales, no resulta fácil de captar. Para ello se requiere la colaboración internacional de diferentes instituciones públicas y privadas que permitan construir modelos representativos de las comunidades de hablantes en sus variedades lingüísticas (geográficas, sociales y estilísticas). Sin un volumen adecuado de datos para entrenar los MLM, estos no podrían generar textos que abarquen y respondan a las características de las distintas variedades del español (Grandury, 2024) o dominios específicos (ámbitos del saber) (Gómez-Pérez, 2023, p. 64). La recopilación de materiales susceptibles de utilizarse en los conjuntos de datos o *datasets,* tanto desde proyectos públicos como privados, sigue siendo insuficiente, pero refleja una creciente concienciación sobre la necesidad de modelos lingüísticos inclusivos en el ámbito hispánico. Los principales retos para la IA en español en los próximos años pueden concretarse en tres aspectos: 1) suficientes datos procedentes de las principales variedades lingüísticas; 2) datos abiertos y disponibles para la investigación y el uso comercial; y 3) datos con calidad suficiente para el entrenamiento de modelos de IA.

La configuración y el volumen de los datos reunidos en los *datasets* son factores cruciales para los MLM. Así, el Gobierno chileno (observa.minciencia.gob.cl) y el Gobierno español (datos.gob.es) han creado portales de datos abiertos que están disponibles para la construcción de estos corpus. Iniciativas como Mozilla Common Voice recopilan voces en diversas lenguas, incluyendo español y catalán, para formar *datasets* con muestras orales, gracias a la participación voluntaria de miles de personas (Mozilla, 2024). Dentro de ILENIA, el proyecto AINA está desarrollando un corpus en catalán con el compromiso de también incluir variedades de esta lengua con pocos recursos digitales, como la aragonesa, con el fin de preservar el patrimonio lingüístico más allá de las lenguas mayoritarias (Barcelona Supercomputing Center, 2024a).

Asimismo, los proyectos NEL-VIVES, NEL-GAITU y NEL-NÓS están creando y ampliando *datasets* con muestras textuales y orales de las lenguas cooficiales del territorio español. El Proxecto Nós, iniciativa de la Universidad de Santiago de Compostela, ha desarrollado un nuevo corpus en gallego que alcanza los 13.95 GB de texto (2.1 mil millones de palabras) para entrenar MLM, denominado *CorpusNÓS* (De-Dios-Flores *et al*., 2024). El proyecto NEL-Vives ha promovido la recopilación de voces en valenciano para formar un corpus que incluya variedades como el alacantí, el tortosí, el meridional, el central y el septentrional (Proyecto NEL-Vives, 2024), además de disponer de varios corpus textuales de carácter administrativo del valenciano (DOGV, BOVA, Les Cortes).

Con una visión global e inclusiva del español, surge el *Corpus del español de negocios CORPEN-FUNDACIÓN COMILLAS*, impulsado por la Fundación Comillas. Este proyecto puede ayudar a suplir las deficiencias de los MLM utilizados en el mundo empresarial, que generalmente están entrenados en inglés. Para 2025, se espera contar con un corpus del español de negocios que incluya datos tanto de España como de Hispanoamérica (Fundación Comillas, 2023). Este corpus complementará el corpus multilingüe de economía y negocios COMENEGO del grupo investigador FRASYTRAM. Por otro lado, el proyecto TeresIA está desarrollando un corpus de terminología científica panhispánica que será accesible a través de un portal con un buscador de metadatos (Gobierno de España, 2023). Este corpus se centrará en la literatura científica producida en España y América Latina, así como en las lenguas cooficiales del territorio español. Además, se busca generar un conjunto de datos terminológicos abiertos (Asociación Española de Terminología, 2024).

Finalmente, en varios países latinoamericanos, se han iniciado proyectos para recopilar datos de lenguas minoritarias con poca o ninguna representación digital. Un ejemplo es la comunidad digital ELOTL en México, que ha creado un corpus paralelo de náhuatl y español (Gutiérrez-Vasques, 2022), así como de mixteco y español, y otomí y español (Comunidad ELOTL, 2020), todos disponibles para descarga y uso en *datasets*. Zeballos *et al*. (2022) desarrollaron un corpus monolingüe de quechua sureño para aprendizaje profundo, como parte del proyecto Llamacha, que sirvió para crear QuBERT, el mayor MLM entrenado para el quechua. Estos esfuerzos reflejan un claro interés en la diversidad hispánica, integrando no solo el español y sus variedades, sino también lenguas minoritarias e incluso en peligro de extinción. Sin embargo, la disponibilidad de *datasets* que capturen esta diversidad lingüística sigue siendo limitada, dependiendo en gran medida de iniciativas de colaboración ciudadana. En muchos casos, estos datos carecen de la calidad y variedad necesarias en cuanto a géneros textuales u otros factores de variación, lo cual representa uno de los principales desafíos para el entrenamiento de los MLM.

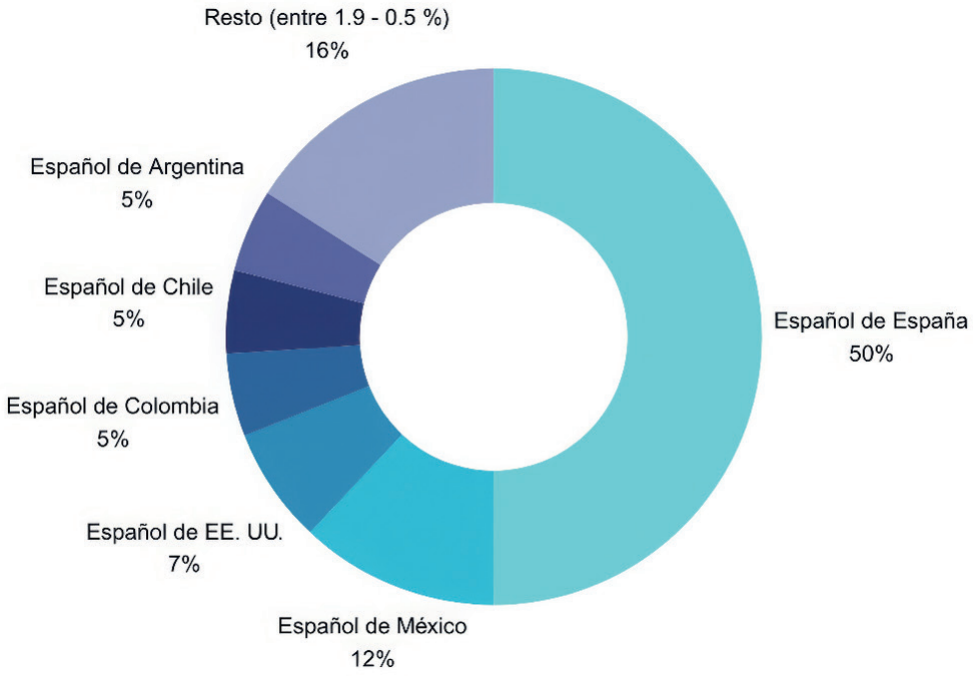
4. Configuración de los MLM para el español

Desde que en 2021 el Centro Nacional de Supercomputación de Barcelona (CNS)/Barcelona Supercomputing Center (BSC) lanzara el proyecto MarIA, se han llevado a cabo varios proyectos de entrenamiento de MLM en español. MarIA consistía en una familia de MLM preentrenados con un enorme corpus del español (2009-2019) obtenido del archivo de la web española de la Biblioteca Nacional de España (BNE). Esta colección, no disponible en abierto, se compone de sitios web españoles (videos, foros, imágenes, documentos, blogs, entre otros) bajo el dominio ".es" y subdominios genéricos como ".com", ".edu", ".gob", ".org", ".net", entre otros) recopilados masivamente (Biblioteca Nacional de España, 2024; Gutiérrez-Fandiño *et al*., 2022).

Pese a la valía de esta iniciativa, no parece claro hasta qué punto y de qué modo esta colección de datos incluye variantes del español. Según el *Language Report Spanish* perteneciente al proyecto europeo *European Language Equality Project*, sobre la diversidad de recursos lingüísticos del español, los corpus en español disponibles en febrero de 2022, incluyendo el de la BNE utilizado por el proyecto MarIA, mostraban porcentajes desproporcionados en cuanto a la representación de la diversidad lingüística de los territorios hispanohablantes. Los porcentajes eran los siguientes: 50 % de español de España, 12 % de español de México, 7 % de español de Estados Unidos; el español de Colombia, Argentina y Chile aparecen con un 5 % cada uno, y se incluye entre 1.9 % y 0.5 % del español del resto de los territorios hispanohablantes (Melero *et al*., 2022, p. 8) (Figura 2).

Además, como se puede apreciar en la Figura 2, la inclusión de materiales de algunos dominios geográficos era mínima, cuando no inexistente o sin especificar, lo que significaba que varios dominios no estaban representados. El entrenamiento con los datos de la BNE de los modelos del proyecto MarIA (RoBERTa-base, RoBERTa-large, GPT-2, GPT-large-bne) parece confirmar un claro sesgo lingüístico en la configuración del modelo, que concede clara preferencia al español peninsular. A esta falta de representatividad se suma el significativo porcentaje de lengua inglesa que constituye la base de algunos modelos, con lo que ello implica para el procesamiento del lenguaje y los resultados que arroja.

Figura 2
Distribución dispar de variedades del español en corpus lingüísticos (a partir de Melero et al. 2022)



El resto de las lenguas cooficiales en España se enfrenta también a esta limitación en cuanto a la representatividad de los datos. El euskera cuenta ya con varias versiones iniciales de un MLM basado en los modelos LLaMa de la empresa Meta (Tabla 2). Estos modelos, llamados *LATXA*, son modelos básicos no disponibles para recibir instrucciones en euskera por parte de usuarios reales al estilo de ChatGPT. Por el contrario, se han concebido como un paso inicial en la investigación hacia un MLM para el euskera que pueda integrarse en herramientas como los *chatbots*. Los modelos LATXA iniciales se construyeron utilizando EusCrawl, un corpus en euskera formado por 12.5 millones de documentos y 423 millones de tokens (Artetxe *et al*., 2022). Una versión mejorada de estos modelos fue entrenada con un corpus de alta calidad que forma el *dataset* público más grande disponible hasta la fecha para el euskera (Etxaniz *et al*., 2024).

Otra iniciativa, dentro del Plan ILENIA, ha sido el desarrollo del modelo Ăguila 7B, entrenado en catalán, español e inglés. En este caso, el porcentaje de cada lengua en el *dataset* fue de un 41.38 % español, un 41.79 % catalán y un 16.84 % inglés (Tabla 2). Por su lado, el modelo FLOR 6.3, otro MLM multilingüe entrenado en catalán, español e inglés, se ha construido sobre el modelo multilingüe BLOOM, al contrario que Ăguila 7B, que parte de un modelo monolingüe inglés (Falcon) (Projecte-aina, 2024a, 2024b). El gallego también cuenta con dos nuevos modelos: Carballo-bloom-1.3B, basado en el modelo multilingüe BLOOM 1.7B y FLOR6.3, y Carballo-cerebras-1.3B, basado en el modelo monolingüe Cerebras-GPT-1.3B y un modelo trilingüe análogo. Ambos tuvieron un preentrenamiento continuo en catalán, español e inglés con los modelos trilingües del proyecto AINA y un posterior entrenamiento solo en gallego (Gamallo *et al*., 2024). En el momento de escribir este artículo, no existe un MLM para el valenciano, aunque el modelo FLOR 6.3B contiene *datasets* en valenciano procedentes del proyecto NEL-VIVES, DOGV, BOVA, Les Cortes.

Tabla 2
Ejemplos de MLM dentro del proyecto ILENIA con sus características de configuración

Modelo	Naturaleza (tipo de datos)	Composición (lenguas)	Objetivos	Desarrollador (fecha de consulta-nombre)
Ăguila 7B	*Datasets* para inglés: Wikipedia. *Datasets* para español: biomedical, legal, Wikipedia, Gutenberg, C4_es. *Datasets* para catalán: C4_ca, biomedical, RacoCatalá Noticias/fórum, CaWaC, Wikipedia, Vilaweb.	Inglés (16.84 %), español, (41.38 %) catalán (41.79 %).	- Crear un modelo *open-source* sobre el modelo Falcon para generación de texto y para refinación de tareas específicas.	Consultado: 08/03/24 BSC Modelo lanzado en julio de 2023.
Flor 6.3B	*Datasets* para el catalán: mc4, MaCoCu, CaWac, oscar-2301, RacoCatala Articles, RacoCatala Forums, Tesis, oscar-2201, Wikipedia, Nació Digital, colossal_oscar_05-06-23, colossal_oscar_03-04-23, colossa_oscar_2022-27, Crawling populars, El Món, ACN, DOGV, DOGE, Vilaweb, hplt, Les Corts Valencianes, IB3, BOUA, Parlament, Aquí Berguedá, Wikimedia, Gutenberg. Datasets para el español: BOE, CSIC, BORME, Tesis (TPX), Eurlex, All_bio_corpora, Wikipedia, colossal_oscar_05-06-23, colossal_oscar_03-04-23. *Datasets* para el inglés: EURLEX, Gutenberg, Wikipedia,legal-mc4, colossal_oscar_05-06-23, colossal_oscar_03-04-23.	Catalán (33.39 %), español (32.32 %), inglés (33.29 %).	- Modelo construido sobre BLOOM 7.1B. - Generar texto y entrenamiento para escenarios específicos.	Consultado: 08/03/24 BSC Modelo lanzado en diciembre de 2023.
LATXA Latxa-7b / Latxa-13b / Latxa-70b	Entrenamiento con *datasets*: EusCrawl, Egunkaria, Booktegi, CulturaX, Wikipedia, Colossal OSCAR, HPLT.	Euskera, inglés (porcentajes no especificados).	- Modelos fundacionales para el euskera a partir de Llama-2. Base de entrenamiento de otros modelos.	Consultado: 11/04/24 Centro Vasco de Tecnología de la Lengua Hitz & IXA Research group Modelos lanzados en 2024.

Modelo	Naturaleza (tipo de datos)	Composición (lenguas)	Objetivos	Desarrollador (fecha de consulta-nombre)
Carballo–bloom-1.3B, Carballo-cerebras-1.3B	*Dataset* a partir de CorpusNÓS.	Inglés, español, catalán, gallego (porcentajes no especificados).	- Modelos fundacionales para el gallego, usando como base, modelos monolingües como Cerebras y multilingües como BLOOM, y modelos trilingües del proyecto AINA.	Consultado: 03/07/2024 Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) Modelos lanzados en junio de 2024.

A pesar de la importancia y la diversidad de las iniciativas comentadas, la mayoría de los MLM para el español suelen derivar de modelos monolingües previamente entrenados en inglés o de modelos multilingües que, en su mayoría, han sido igualmente entrenados en inglés. Estos modelos utilizan la técnica de preentrenamiento continuo, lo que significa que, en lugar de entrenar un modelo desde cero, se toma uno ya existente (en este caso, en inglés) al que se le incorporan nuevos datos o se le añade una nueva lengua (Gamallo *et al*., 2024, p. 4) como parte del nuevo entrenamiento.

A principios de 2024, se anunció la creación de un modelo fundacional para el español; es decir, un modelo entrenado con gran cantidad de datos en español no etiquetados, que puede utilizarse para crear otros modelos especializados (*Small Language Models*, *SLMs*) destinados a tareas específicas. Este modelo aspira a ser el primero de su magnitud en el mundo, de código abierto y transparente, entrenado totalmente en español y en las lenguas cooficiales del territorio español (Gallardo, 2024). Este desarrollo se basará en el trabajo ya realizado en el proyecto ILENIA (véase la Tabla 1). Más recientemente, la nueva *Estrategia de inteligencia artificial* (Gobierno de España, 2024) ha redefinido esta idea inicial, proponiendo la creación de una familia de modelos generativos y no generativos, tanto multilingües como monolingües, para el español, conocida como ALIA. En el último cuatrimestre del año, se espera lanzar modelos multilingües (español y lenguas cooficiales en España), mejorando los ya existentes, así como el primer modelo ALIA, que partirá de los modelos en ILENIA (véase la Tabla 1) y que estará entrenado con al menos un 20 % de datos en español (Alonso, 2024). A partir de 2025, se aspira a desarrollar modelos monolingües para el español de hasta 175B. Los modelos ALIA serán los primeros de su magnitud para el español, de código abierto y transparente (Gallardo, 2024).

El elevado coste económico y la cantidad de recursos necesarios para entrenar un modelo fundacional en español contrasta con las iniciativas llevadas a cabo por pequeñas comunidades de expertos en aprendizaje automático, lingüística computacional y procesamiento del lenguaje natural. Un buen ejemplo es la comunidad #somosPLN, que promueve la democratización de los MLM en español organizando *hackathons*, encuentros donde programadores, desarrolladores y *hackers* contribuyen al avance de estos modelos. De uno de estos encuentros surgió el MLM monolingüe en español, BERTIN (BERTIN GPT-J-6B, BERTIN RoBERTa) (De la Rosa *et al*., 2022, p. 14), publicado casi a la par que los modelos del proyecto MarIA en 2022. Ya en 2020, había aparecido el modelo monolingüe BETO (Spanish Pre-Trained BERT model), iniciativa de

académicos de la Universidad de Chile (Cañete *et al*., 2023). Otros MLM impulsados de forma independiente han sido RigoBERTa (Vaca Serrano *et al*., 2022), ALBETO y DistilBETO (Cañete *et al*., 2022), Electricidad, MT0, BioMedtra, LINCE-ZERO (de la empresa privada Clibrain) o LeNIA (Moreno Sandoval, 2024).

Es importante destacar que no existe un panel de clasificación para los MLM en español, como sí existe para los modelos entrenados en inglés, lo que impide realizar una comparación clara de todos los modelos existentes hasta la fecha, evidenciando una vez más la falta de coordinación en el desarrollo de la IA en español. Del mismo modo, las entidades que gestionan los proyectos no suelen compartir sus algoritmos y técnicas de entrenamiento exclusivos o utilizan *datasets* patentados, lo que impide su acceso y disponibilidad para otros grupos interesados en la investigación y el entrenamiento de modelos en español (De la Rosa *et al*., 2022, p. 14). No obstante, algunos proyectos recientes, como AINA, han declarado que sus modelos son de acceso abierto, permitiendo que cualquiera los utilice (Barcelona Supercomputing Center, 2024b).

Por último, se plantean algunos interrogantes sobre el desarrollo de los MLM para el español. En primer lugar, aún no está claro en qué medida los modelos monolingües en español reflejan la diversidad lingüística del idioma. Y, en segundo lugar, la creación de MLM multilingües, como aquellos entrenados con datos en español y en diversas lenguas cooficiales, aún presenta dudas sobre su rendimiento en comparación con los modelos monolingües y sus habilidades de transferencia lingüística cruzada (Kew *et al*., 2023), así como sobre la importancia relativa de la calidad de los datos frente al porcentaje de datos de entrenamiento en cada lengua.

En definitiva, es evidente que la combinación y coordinación de iniciativas como las mencionadas anteriormente, tanto institucionales como independientes, podrían conducir a una mejora de la IA en español y que en los próximos años podríamos observar resultados positivos. Sin embargo, la falta de un eje común en el ámbito hispanohablante se hace evidente para los MLM en español. Queda por ver qué estrategia se desplegará con el objetivo de reflejar en estos modelos la diversidad lingüística de los territorios hispanohablantes.

5. El Sesgo Lingüístico Digital (SLD): una necesidad de acción coordinada para la IA en español

En la primera sección, se han examinado diversas propuestas nacionales de países hispanohablantes para maximizar el rendimiento de la IA generativa. Estas iniciativas proponen líneas de acción que abarcan temas como la educación, la salud, la participación ciudadana y la democracia, el cambio climático, la biodiversidad, la automatización y la investigación académica (véase VV. AA., 2023, p. 162). No obstante, una de las grandes incógnitas que surge en toda esta información es cuáles de los ámbitos mencionados se considerarán prioritarios según el contexto geográfico y cómo cada uno de los países hispanohablantes impulsará la creación de MLM. Entre todas estas prioridades, debido a la importancia de la lengua como parte integral de la IA generativa —por la variedad de fuentes lingüísticas y textuales que componen estos modelos— la monitorización del material lingüístico que nutra la creación de estos MLM debería ser un aspecto central en las hojas de ruta nacionales y en la acción coordinada de los países hispanohablantes. No implementar este tipo de acción en el diseño de los MLM podría conducir a importantes sesgos en la representatividad del material lingüístico. Como señalan Helm *et al*. (2024, p. 8):

> Language modelling bias is observed when the technology, by design, represents, interprets, or processes content less accurately in certain languages than in others, thereby forcing speakers of the disadvantaged language to simplify or adapt their communication, (self-)representation, and expression when using that technology to fit the default incorporated in the privileged language.

Helm *et al*. (2024) se centran en el nivel interlingüístico; es decir, en cómo se desarrollan los MLM para lenguas más o menos cercanas o distantes del inglés. Sin embargo, para el caso específico del español, es necesario considerar, desde un nivel intralingüístico, las distintas variedades de la lengua. Para ello, proponemos el término *Sesgo Lingüístico Digital* (*SLD*) o *Digital Linguistic Bias* (*DLB*) para referirnos a la hibridez lingüística que la tecnología genera tanto a nivel interlingüístico (p. ej., en relación con la base del inglés utilizada para entrenar estos modelos) como intralingüístico (en relación con las distintas variedades de la lengua). Esta combinación puede dar lugar a diferentes sesgos, desde la sobreabundancia de calcos lingüísticos del inglés, tanto a nivel léxico como sintáctico (nivel interlingüístico), hasta una presencia mayoritaria o desproporcionada de las variedades dominantes o hegemónicas, frente a otras variedades menos representadas en el material que constituye la base del modelo (nivel intralingüístico).

Esta hibridez lingüística (véase Kabatek, 2011, pp. 271-272, sobre diferentes aproximaciones a este término) se obtiene al superponer ambos niveles, el inter- y el intralingüístico, alejando de lo autóctono, real u original el material lingüístico que proporciona un *chatbot* y dando lugar, en los MLM, a combinaciones indiscriminadas de rasgos procedentes de distintas variedades (a nivel fonético-fonológico, morfosintáctico, léxico e incluso pragmático). Las consecuencias de esta configuración de los MLM se observan en distintos planos. En primer lugar, dada la accesibilidad de las herramientas potenciadas por IA, la obtención de material lingüístico sesgado puede tener un impacto en los usos lingüísticos de los hablantes y en el conjunto del patrimonio lingüístico hispanohablante, al poder asumirse los resultados como modelo de lengua. En segundo lugar, es también importante considerar la representatividad de los MLM en relación con los diferentes modos de comunicación, tanto en la oralidad (p. ej., la dialogicidad), como en la escritura (p. ej., los rasgos prototípicos de un género textual) (Del Rey Quesada, 2021). Y, en tercer lugar, en el caso de los *chatbots*, la calidad de las instrucciones o *prompts* utilizados (un *prompt* general vs. un *prompt* pautado y con coordenadas específicas sobre la realización de la tarea) puede llevar también a la aparición de sesgos.

Actualmente, los MLM en español no contemplan una diversidad de modelos que reflejen diferentes variedades de la lengua. De modo similar a cualquier corpus lingüístico general y no especializado, en los MLM predomina el material lingüístico procedente de zonas con un mayor peso demográfico. Sin embargo, la diferencia radica en que, mientras que un corpus sirve de herramienta para la consulta de los datos almacenados, un MLM, como modelo probabilístico de base estadística, es capaz de generar textos y contenidos, sirviéndose de múltiples muestras de lengua sin prestar atención a su composición dialectal o, en general, lingüística. Como mencionamos arriba, el usuario interacciona con la herramienta (p. ej. *chatbot*) a partir de instrucciones o *prompts* y la herramienta es capaz de producir enunciados gramaticalmente aceptables. Del mismo modo, el texto generado refleja propiedades textuales como la coherencia, en la disposición lógica de los párrafos (introducción, desarrollo y conclusión), y la cohesión o relación interna de la información y de los párrafos, por ejemplo, mediante el uso de marcadores discursivos o frases hechas. Sin

embargo, todo esto se hace sin importar la adscripción a una variedad dialectal específica ni a un contexto sociocultural determinado. En el caso de los *chatbots*, como ChatGPT, la capacidad de generar textos coherentes o cohesionados hace que el usuario los perciba como completos y correctos y, por lo tanto, al mismo nivel que los textos generados por humanos, aunque se perciba en ellos una artificialidad que a menudo no se sabe explicar.

Del mismo modo, se pueden identificar al menos cuatro aspectos fundamentales que influyen directamente en la aparición de sesgos lingüísticos digitales (Gómez-Pérez, 2024, pp. 82-83): 1) las alucinaciones (respuestas incompletas o incorrectas debido a datos de entrenamiento inexactos, desactualizados o no disponibles), 2) la necesidad de ajuste fino (datos actualizados y anotados por humanos), 3) la calidad y cantidad de los datos (los modelos solo pueden responder a preguntas específicas del dominio para el que han sido entrenados, lo que es un problema para dominios específicos que carecen de corpus extensos, así como para lenguas o variedades minoritarias con recursos limitados), y 4) la falta de trazabilidad (no mantener la referencia al documento original de la información, lo que genera desconfianza y dificulta discernir si la información proviene de los datos de entrenamiento o es una suposición estadística del modelo).

Llegados a este punto, planteamos la necesidad de una acción coordinada para los MLM en español que refleje y preserve la diversidad, la complejidad y las particularidades de la lengua, aspecto tan defendido por los *agentes* (institucionales) y las *voces* (docentes e investigadores) del español (Muñoz-Basols y Hernández Muñoz, 2019, p. 82). De no hacerse así, la IA, a través de los MLM, podría tener un impacto negativo en las llamadas tres *PPP* del español: el español como *lengua policéntrica* (proyectada desde diferentes polos en sus distintas variedades; es decir, construida sobre una norma culta policéntrica) (Moreno Fernández, 2000), el español como *lengua polifónica* (ecosistema de variedades que conecta a sus hablantes) y como *lengua poliédrica* (con las múltiples caras o perfiles de sus hablantes: lengua oficial, herencia generacional o lengua de aprendizaje, entre otros) (Muñoz-Basols y Hernández Muñoz, 2019, pp. 81-82).

Como parte de esta acción coordinada en busca de una adecuada representatividad, sería necesario abordar de un modo apropiado la recopilación masiva de fuentes escritas y orales. Con todo, este plan de acción para las tecnologías de la lengua debería mostrar una "conciencia digital" que busque desarrollar estrategias y políticas que preserven la diversidad de la lengua en el medio digital, tal y como se propugna desde el Parlamento Europeo:

> El Parlamento Europeo insta a la Comisión Europea y a los Estados miembros a desarrollar estrategias y acciones políticas para facilitar la diversidad lingüística en el mercado digital y, en ese contexto, insta a la Comisión y a los Estados miembros a determinar los recursos lingüísticos mínimos que deben tener todas las lenguas europeas para evitar su extinción digital, como colecciones de datos, léxicos, grabaciones orales, memorias de traducción, corpus etiquetados y contenidos enciclopédicos. (Gobierno Vasco, 2021, p. 16; traducido del inglés)

De este modo, el programa GAITU, desarrollado por el Gobierno Vasco para el euskera, es un ejemplo de esta estrategia con "su puesta a disposición de las empresas, personas investigadoras y demás personas interesadas del sector" (Gobierno Vasco, 2021, p. 3). En el plano de la oralidad, destaca el diseño de interfaces de voz adecuadas para obtener voces naturales y espontáneas, que

expresen personalidad y emociones, para lo que es necesario incluir voces diversas y expresivas, teniendo en cuenta diferentes tonos, estilos y dialectos (Gobierno Vasco, 2021, p. 16). En este ámbito, posiblemente sean las lenguas con menor número de hablantes las más concienciadas de la importancia de esta labor y las que consideran más necesaria la puesta en marcha de una acción coordinada que salvaguarde su presencia y representatividad en el ámbito digital.

Más allá de los ya conocidos ideologemas en torno a la lengua como "el español es una lengua de encuentro y diálogo", "el español es una lengua universal", "el español, lengua de todos" (Moreno Fernández, 2019), debe existir un compromiso real por preservar la diversidad del idioma en una acción conjunta y coordinada en torno a la IA. Si no se logra una adecuada representatividad de las variedades del español en los MLM, podría estar promoviéndose una pseudolengua, una lengua neutral artificial, así como una fragmentación que llevaría a la formación de "dialectos digitales", como lo describe el director de la RAE, Santiago Muñoz Machado (*El Debate*, 2024), comprensibles, pero no representativos de comunidades de hablantes reales. En el VII Congreso Internacional de la Lengua Española (CILE) de San Juan (Puerto Rico), Francisco Moreno Fernández recordaba la necesidad de una correspondencia del ámbito digital con la realidad policéntrica y multinormativa de la lengua española, que habría de recibir un tratamiento prioritario:

> La lengua española debe ser habilitada para todas las innovaciones tecnológicas que se vayan produciendo, haciendo posible que, por ejemplo, todos los protocolos, aplicaciones y recursos técnicos desplegados para la comunicación automatizada, la transmisión de información y las redes sociales acepten las peculiaridades formales del español. (Moreno Fernández, 2016, en línea)

Para ello, es necesario alcanzar en los MLM una alta representatividad del material lingüístico en español, hecho que requiere una acción coordinada y conjunta de los diferentes territorios y agentes institucionales hispanohablantes.

6. Limitaciones, futuras líneas de investigación y conclusiones

Este trabajo ha puesto en evidencia algunas de las limitaciones a las que se enfrenta la IA en español, entre las que no es menor la imposibilidad de realizar un análisis detallado de la composición de los MLM existentes. Así, en la Tabla 1 y la Tabla 2, no ha sido posible incluir para todos los MLM el porcentaje de los componentes lingüísticos que integran cada modelo. Esto se debe a la opacidad de los propios MLM y a que el acceso a este tipo de información no resulta ni fácil ni previsible. Además, no hay información disponible sobre el volumen y la proporción de las diferentes variedades geolectales, campos temáticos, estilos, géneros textuales, entre otros aspectos, incluidos en los MLM. Esto afecta al procesamiento del lenguaje, dada la diversidad de medios digitales compilados, con características propias, tanto de la oralidad (p. ej., redes sociales) como de la escritura (p. ej., artículos académicos, libros), que no se discriminan adecuadamente. Todos los textos se generan de forma completamente descontextualizada, por lo que su calidad se fundamenta en un carácter neutro o neutralizado y, por lo tanto, alejado de contextos de los usos reales, vinculados a diferentes geografías y espacios sociales.

Como consecuencia de todo ello, algunas líneas de investigación futuras podrían enfocarse en encontrar mecanismos que optimicen y acerquen el uso de la IA a nivel de usuario desde una

perspectiva lingüística. Sería importante que los materiales lingüísticos con los que se entrenan los MLM incluyeran marcas contextualizadoras y relativas a usos variables de la lengua, de la misma manera que sería beneficioso desarrollar *chatbots* que permitieran a los usuarios seleccionar la variedad de lengua con la que se identifican, similar a un selector de idioma en un procesador de texto o utilizar la geolocalización del usuario para proponer opciones pertinentes y mejor contextualizadas para comunidades de hablantes o variedades específicas. Asimismo, podría crearse una función que discrimine y compare características geolectales, sociolectales, textuales, estilísticas diferentes, lo que no solo sería útil para entrenar los MLM, sino también para su aplicación en otros ámbitos, como la enseñanza de la lengua (Muñoz-Basols *et al*., 2023; Muñoz-Basols y Fuertes Gutiérrez, 2024). No obstante, como afirma García Montero (2022), "enseñar español a las máquinas y que estas nos ayuden a enseñarlo es el reto lingüístico y cultural más importante del siglo XXI" (p. 140).

Como hemos visto, los datos lingüísticos que alimentan a la IA y que continuarán nutriéndola en los próximos años serán cruciales para asegurar la calidad de los resultados obtenidos. Además de tomar las medidas necesarias para mitigar los efectos del Sesgo Lingüístico Digital, tal como se ha discutido en este artículo, es esencial evitar caer en el aforismo GIGO (*Garbage in, Garbage out*), que refleja cómo la calidad del resultado (*output*) de un sistema de IA está directamente relacionada con la calidad de los datos de entrada (*input*) (Roden *et al*., 2022). En otras palabras, al crear un MLM, es imprescindible garantizar que los datos de entrada sean de alta calidad para obtener resultados igualmente válidos.

Además de la necesidad de monitorizar qué datos se utilizan para entrenar los MLM en español, es importante destacar la escasez de estos, como la disponibilidad limitada de textos generados por humanos. De ahí que la colaboración y acción coordinada entre los agentes institucionales de los países hispanohablantes, mediante la digitalización de sus repertorios lingüísticos, cobre aún más relevancia para asegurar el desarrollo adecuado de la IA en español. Aunque la presencia digital del español ha seguido creciendo y la accesibilidad a recursos digitales ha mejorado, la falta de nuevos datos en forma de textos puede convertirse en un "auténtico muro" (*data wall*) que impida el avance y entrenamiento continuo de los MLM en español. Las herramientas de IA consumen datos a un ritmo más rápido del que se producen, ya sea procedentes de textos, bases de datos, corpus o libros disponibles en internet.

Se estima que, si las tendencias actuales en el desarrollo de los MLM continúan, los modelos se entrenarán con conjuntos de datos que tendrán un tamaño similar al total de datos públicos de texto humano disponible entre 2026 y 2032, o incluso antes, en 2028, si se produce un sobreentrenamiento (Villalobos *et al*., 2022). Esto podría llevar a la necesidad de generar datos sintéticos, es decir, creados artificialmente y no recopilados de la realidad humana (Villalobos *et al*., 2022). En el caso del español, queda claro que será fundamental continuar creando repositorios de datos y textos representativos del mundo hispanohablante actual y, por lo tanto diversos, y con la calidad suficiente para entrenar eficazmente los MLM en español.

Durante el Mobile World Congress 2024, se anunció la creación de un gran modelo fundacional de lenguaje de inteligencia artificial, en colaboración con el Barcelona Supercomputing Center, la Red Española de Supercomputación, la RAE y la ASALE (Portal Administración Electrónica, 2024). No obstante, como veíamos al comienzo de este artículo, la incorporación de la IA generativa a nivel de usuario ha dado lugar a una diversidad de propuestas desde los países hispanohablantes enfocadas

en el desarrollo de iniciativas independientes que no muestran una "estrategia global" para la lengua. De este modo, el análisis de la situación actual y de cómo se está abordando esta tecnología en el ámbito hispanohablante revelan una desconexión entre los distintos agentes institucionales y proyectos. Por ello, es fundamental que se aborden trabajos colaborativos y transdisciplinarios, que prioricen la integración de aspectos tanto tecnológicos como lingüísticos en los MLM para el español.

Para posibilitar el desarrollo y la mejora de los MLM en la calidad y representatividad del material lingüístico que los conforma, es necesario que los equipos de trabajo cuenten con el asesoramiento de expertos en cuestiones de variedades y usos lingüísticos. Estos especialistas son conocedores de la materia prima que alimenta, sustenta y entrena estos modelos. Si no se crea un modelo fundacional del español a partir de una colaboración entre los países hispanohablantes, será más difícil obtener un resultado suficientemente representativo de la diversidad hispánica. Hasta ese momento, un aspecto fundamental de la IA en español consistirá en concienciar a la sociedad sobre la existencia de un claro Sesgo Lingüístico Digital y sobre la necesidad de buscar mecanismos que lo eviten y mitiguen.

Referencias

Agesic. (2019). *Estrategia de Inteligencia Artificial para el Gobierno Digital*. Gobierno de Uruguay.

Alonso, R. (2024, 15 de mayo). El gobierno acelera el desarrollo de ChatGPT español y el uso de la Inteligencia Artificial en pymes. *ABC*. https://www.abc.es/tecnologia/gobierno-destinara-1500-millones-euros-desarrollo-ia-20240514132352-nt.html?ref=https%3A%2F%2Fwww.google.com%2F

Amaratunga, T. (2023). *Understanding Large Language Models*. Apress.

Arancibia, D., Ávila, C., Caro, M. J., Girardi, J., González, N., Guridi, J. A. y Rivera, A. (2021). *Política Nacional de Inteligencia Artificial. Chile IA*. Ministerio de Ciencia, Tecnología, Conocimiento e Innovación.

Artetxe, M., Aldabe, I., Agerri, R., Perez-de-Viñaspre, O. y Soroa, A. (2022). *Does corpus quality really matter for low-resource languages?* arXiv. https://doi.org/10.48550/arXiv.2203.08111

Asociación Española de Terminología. (2024). *TERESIA*. https://aeter.org/teresia/

Biblioteca Nacional de España. (2024). *El Archivo de la Web España*.https://www.bne.es/es/colecciones/archivo-web-espanola

Barcelona Supercomputing Center. (2024a, 17 de enero). *BSC to develop multilingual models in Aranese through Aina* [nota de prensa]. https://www.bsc.es/news/bsc-news/bsc-develop-multilingual-models-aranese-through-aina

Barcelona Supercomputing Center. (2024b, 28 de febrero). *El BSC pone en marcha Aina Challenge, la primera convocatoria oficial de proyectos de inteligencia artificial en catalán* [nota de prensa]. https://www.bsc.es/es/noticias/noticias-del-bsc/el-bsc-pone-en-marcha-aina-challenge-la-primera-convocatoria-oficial-de-proyectos-de-inteligencia

Campusa. (2024, 31 de enero). HiTZ Zentroa desarrolla el mayor modelo del lenguaje para el euskera: Latxa. *Campusa, Noticias de la Universidad del País Vasco*. https://www.ehu.eus/es/-/hitz-zentroa-desarrolla-mayor-modelo-lenguaje-euskera-latxa

Cañete, J., Chaperon, G., Fuentes, R., Ho, J., Kang, H. y Pérez, J. (2023). Spanish Pre-trained BERT Model and Evaluation Data. *arXiv*. https://doi.org/10.48550/arXiv.2308.02976

Cañete, J., Donoso, S., Bravo-Marquez, F., Carvallo, A. y Araujo, V. (2022). ALBETO and DistilBETO: Lightweight Spanish Language Model. *arXiv*.

Comunidad ELOTL (2020). *Corpus paralelo Otomí-español*. https://elotl.mx/proyectos/corpus-paralelo-otomi-espanol/

Company Company, C. (2019). Jerarquías dialectales y conflictos entre teoría y práctica. Perspectivas desde la Asociación de Academias de la Lengua Española (ASALE). *Journal of Spanish Language Teaching*, *6*(2), 96-105. https://doi.org/10.1080/23247797.2019.1668179

Dafoe, A. (2018). *AI Governance: a Research Agenda*. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf

De-Dios-Flores, I., Paniagua Suárez, S., Carbajal Pérez, C., Bardanca Outeiriño, D., Garcia, M. y Gamallo, P. (2024). *CorpusNÓS: A massive Galician corpus for training large language models*. arXiv. https://iv.org/html/2406.13893v1/arx

De la Rosa, J., Ponferrada, E. G., Villegas, P., González, P., Romero, M. y Grandury, M. (2022). BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Procesamiento del Lenguaje Natural*, *68*, 13-23. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403

Del Rey Quesada, S. (2021). Lo marcado y lo no marcado en la cadena de variedades: apuntes para una nueva propuesta. En T. Gruber, K. Grübl y T. Scharinger (Eds), *Was bleibt von kommunikativer Nähe und Distanz? Mediale und konzeptionelle Aspekte sprachlicher Variation* (pp. 205-238). Narr.

El Debate. (2024, 30 de mayo). Santiago Muñoz Machado: "El peligro está en que se formen dialectos digitales que laminen nuestro idioma". *El Debate*. https://www.eldebate.com/sociedad/20240530/santiago-munoz-machado-peligro-esta-formen-dialectos-digitales-laminen-nuestro-idioma_201254.html

Etxaniz, J., Sainz, O., Perez, N., Aldabe, I., Rigau, G., Aguirre, E., Ormazabal, A., Artetxe, M. y Soroa, A., (2024). Latxa: An Open Language Model and Evaluation Suite for Basque. *arXiv*. https://doi.org/10.48550/arXiv.2403.20266

Fundación Comillas. (2023). *Corpus del español de los negocios (CORPEN-FUNDACIÓN COMILLAS)*. https://fundacioncomillas.es/wp-content/uploads/2023/03/proyecto-corpen-fundacion-comillas.pdf

Gallardo, C. (2024, 28 de febrero). Spain to develop open-source LLM trained in Spanish, regional languages. *Sifted*. https://sifted.eu/articles/spain-large-language-model-generative-ai

Gamallo, P., Rodríguez, P., de-Dios-Flores, I., Sotelo, S., Paniagua, S., Bardanca, D., Pichel, J. R. y Garcia, M. (2024). Open Generative Large Language Models for Galician. ArXiv. https://arxiv.org/pdf/2406.13893v1

García Montero, L. (2022). Reflexiones precavidas sobre la inteligencia artificial. En C. Pastor Villalba (Dir.), *El español en el mundo 2022. Anuario del Instituto Cervantes* (pp. 135-144). Instituto Cervantes.

Gobierno de España. (2022, 26 de junio). *El gobierno concede una subvención de 5 millones de euros a la RAE para ejecutar el proyecto 'Lengua española e Inteligencia Artificial' (LEIA)*. https://planderecuperacion.gob.es/noticias/el-gobierno-concede-una-subvencion-de-5-millones-de-euros-a-la-rae-para-ejecutar-el-proyecto-leia

Gobierno de España (2023, 15 de diciembre). *Conoce "TeresIA" para la traducción de terminología en español mediante Inteligencia Artificial*. https://planderecuperacion.gob.es/noticias/conoce-proyecto-teresia-traduccion-terminologia-espanol-inteligencia-artificial-IA-prtr

Gobierno de España (2024). *Estrategia de Inteligencia Artificial 2024*. https://portal.mineco.gob.es/es-es/digitalizacionIA/Documents/Estrategia_IA_2024.pdf

Gobierno Vasco. (2021). *Plan de Acción de Las Tecnologías de La Lengua 2021-2024*. Departamento de Cultura y Política Lingüística.

Gómez-Pérez, A. (2023). *Inteligencia artificial y lengua española* [Discurso de ingreso]. Real Academia Española.

Gómez-Pérez, A. (2024). *Ingeniería ontológica* [Discurso de ingreso]. Real Academia de Ingeniería.

Grandury, M. [@SomosPLN]. (2024, 13 de marzo). *Diversidad lingüística e IA, cómo desarrollar LLMs inclusivos* [Vídeo]. Youtube. https://www.youtube.com/watch?v=QCNPVy3QWFs

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C. , Rodríguez-Penagos y Villegas, M. (2022). Maria: Spanish language models. *ArXiv*. https://doi.org/10.48550/arXiv.2107.07253

Gutiérrez-Vasques, X. [@SomosPLN]. (2022, 30 de marzo). *Consideraciones de NLP para lenguas minorizadas. El caso de México. Hackathon de PLN en español.* [Vídeo]. Youtube. https://www.youtube.com/live/aNR7UM-E6vA?si=H1LnC7F6jqFlA_el

Helm, P., Bella, G., Koch, G. y Giunchiglia, F. (2024). Diversity and language technology: how language modeling bias causes epistemic injustice. *Ethics Inf Technol*, *26*, 1-8. https://doi.org/10.1007/s10676-023-09742-6

Instituto Cervantes y Ministerio de Economía y Transformación Digital (2023). *Estado actual de los corpus en español, lenguas cooficiales y variantes del español*. Instituto Cervantes y Ministerio de Economía y Transformación Digital.

Impulso de las Lenguas en la Inteligencia Artificial. (2024). *Sobre Ilenia*. https://proyectoilenia.es/sobre-ilenia/

Kabatek, J. (2011). Algunos apuntes acerca de la cuestión de la "hibridez" y de la "dignidad" de las lenguas iberorrománicas. En Y. Congosto y E. Méndez (Coords.), *Variación lingüística y contacto de lenguas en el mundo hispánico: in memoriam Manuel Alvar* (pp. 271-289). Iberoamericana.

Kew, T., Schottmann, F. y Sennrich, R. (2023). Turning English-centric LLMs Into Polyglots: How much Multilinguality is needed? *ArXiv*. https://doi.org/10.48550/arXiv.2312.12683

Lagunes A., Martínez Y., Cárdenas C., De la Peña S., Mancilla D., Xilotl R., Sánchez O., Moguel A. y Cárdenas J. (2024). *Propuesta de Agenda Nacional de la Inteligencia Artificial para México (2024 - 2030)*. Alianza Nacional de Inteligencia Artificial (ANIA).

Liu, Y., Cao, J., Liu, C., Ding, K. y Jin, L. (2024). Datasets for Large Language Models: A comprehensive Survey. *arXiv*. https://arxiv.org/pdf/2402.18041

Marres, N. (2017). *Digital sociology: The reinvention of social research*. Polity Press.

Melero, M., Peñarrubia, P., Cabestany, D., Figueras, B. C., Rodríguez, M. y Villegas, M. (2022). *D1.32 Report on the Spanish Language*. European Language Equality.

MinCiencia. (2024). *Política Nacional de Inteligencia Artificial*. Gobierno de Chile.

Ministerio de Ciencia, Tecnología e Innovación. (2024). *Hoja de ruta para el desarrollo y aplicación de la Inteligencia Artificial en Colombia*. Dirección de Desarrollo Tecnológico e Innovación.

Moreno Fernández, F. (2000). *Qué español enseñar*. Arco/Libros.

Moreno Fernández, F. (2016). *La búsqueda de un 'español global'* [Ponencia]. VII Congreso Internacional de la Lengua Española. Instituto Cervantes, Real Academia Española y Asociación de Academias de la Lengua Española. https://congresosdelalengua.es/puerto-rico/paneles-ponencias/espanol-mundo/moreno-fancisco.htm

Moreno Fernández, F. (2019). El español en movimiento. En F. Moreno Fernández (Coord.), *Archiletras Científica 2. El español, lengua migratoria* (pp. 20-25). Prensa y Servicios de la Lengua SLU.

Moreno Fernández, F. (2022). La variación geográfica y social en los corpus lingüísticos. En G. Parodi, P. Cantos-Gómez y Ch. Howe (Eds.), *Lingüística de corpus en español. The Routledge Handbook of Spanish Corpus Linguistics* (pp. 296-309). Routledge.

Moreno Fernández, F. y Cestero Mancera, A. M. (2020). El proyecto PRESEEA: desarrollos analíticos. *Verba*: *Anuario Galego de Filoloxía*, *80*, 119-138. https://dx.doi.org/10.15304/9788418445316

Moreno Sandoval, A. (2024). *El español artificial. El español en el mundo. Anuario del Instituto Cervantes*. Instituto Cervantes.

Mozilla. (2024). Common Voice datasets. https://commonvoice.mozilla.org/en/datasets

Muñoz-Basols, J., Craig, N., Lafford, B. A. y Godev, C. (2023). Potentialities of Applied Translation for Language Learning in the Era of Artificial Intelligence. *Hispania*, *106*(2), 171–194. https://doi.org/10.1353/hpn.2023.a899427

Muñoz-Basols, J. y Fuertes Gutiérrez, M. (2024). Oportunidades de la Inteligencia Artificial (IA) en la enseñanza y el aprendizaje de lenguas. En J. Muñoz-Basols, M. Fuertes Gutiérrez y L. Cerezo (Eds.), *La enseñanza del español mediada por tecnología: de la justicia social a la Inteligencia Artificial (IA)* (pp. 343-364). Routledge. https://doi.org/10.4324/9781003146391-18

Muñoz-Basols, J. y Hernández Muñoz, N. (2019). El español en la era global: agentes y voces de la polifonía panhispánica. *Journal of Spanish Language Teaching, 6*(2), 79-95. https://doi.org/10.1080/23247797.2020.1752019

Nguyen, D. y Hekman, E. (2022). The news framing of artificial intelligence: A critical exploration of how media discourses make sense of automation. *AI & SOCIETY*, *39*, 437-451. https://doi.org/10.1007/s00146-022-01511-1

Peláez Agudo, D. (2023). *El impacto de la revolución de la IA en España y Latinoamérica.* OBS Business School.

Portal Administración Electrónica. (2024, 27 de febrero). *El Gobierno anuncia la construcción de un modelo de lenguaje de IA entrenado en español y las lenguas cooficiales*. https://administracionelectronica.gob.es/pae_Home/pae_Actualidad/pae_Noticias/2024/Febrero/Noticia-2024-02-27-Gobierno-anuncia-modelo-fundacional-lenguaje-IA.html

Presidencia de la Nación. (2020). *Plan Nacional de Inteligencia Artificial*. Gobierno de Argentina.

Presidencia del Consejo de Ministros. (2021). *Estrategia Nacional de Inteligencia Artificial. Documento de Trabajo para la participación de la ciudadanía 2021-2026*. Secretaría de Gobierno y Transformación Digital.

Projecte-aina. (2024a, 21 de julio). *FLOR-6.3B*. Hugging Face. https://huggingface.co/projecte-aina/FLOR-6.3B

Projecte-aina. (2024b, 21 de julio). *Ǎguila-7B*. Hugging Face. https://huggingface.co/projecte-aina/aguila-7b

Proyecto NEL-Vives. (2024, 21 de julio). *How to give your voice.* VIVES. https://vives.gplsi.es/instruccions/

Real Academia Española. (2022a, 20 de mayo). *El presidente de Microsoft visita la RAE* [Nota de prensa]. Real Academia Española. https://www.rae.es/noticia/el-presidente-de-microsoft-visita-la-rae

Real Academia Española. (2022b, 26 de mayo). *La RAE y AWS presentan una herramienta basada en inteligencia artificial para conocer el estado del español en Internet* [Nota de prensa]. Real Academia Española. https://www.rae.es/noticia/la-rae-y-aws-presentan-una-herramienta-basada-en-inteligencia-artificial-para-conocer-el

Real Academia Española y Asociación de Academias de la Lengua Española. (2004). *La nueva política lingüística panhispánica*. Real Academia Española.

Roden, B., Lusher, D., Spurling, T. H., Simpson, G. W., Klein, T., Brailly, J. y Hogan, B. (2022). Avoiding GIGO: Learnings from data collection in innovation research. *Social Networks*, 69, 3–13. https://doi.org/10.1016/j.socnet.2020.04.005

Vaca Serrano, A., García Subies, G., Montoro Zamorano, H., Aldama García, N., Samy, D., Betancur Sánchez, D., Moreno-Sandoval, A., Guerrero Nieto, M. y Barbero Jiménez, Á. (2022). Rigoberta: a state-of-the-art language model for Spanish. *ArXiv*. https://doi.org/10.48550/arXiv.2205.10233

Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M. y Ho, A. (2022). Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv*. https://doi.org/10.48550/arXiv.2211.04325

VV. AA. (2023). *Índice latinoamericano de inteligencia artificial. Centro Nacional de Inteligencia Artificial*. CENIA.

Zeballos, R., Ortega, J., Chen, W., Castro, R., Bel. N., Yoshikawa, C., Ventura, R., Aradiel, H. y Melgarejo, N. (2022). Introducing QuBERT: A Large Monolingual Corpus and BERT model for Southern Quechua. En C. Cherry, A. Fan, G. Foster, G. Haffari, S. Khadivi, N. Peng, X. Ren, E. Shareghi y S. Swayamdipta (Eds.), *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing* (pp. 1-13). Association for Computational Linguistics.

Contribución de los autores

Javier Muñoz-Basols diseñó y planificó el desarrollo del trabajo de investigación; Javier Muñoz-Basols y María del Mar Palomares Marín han participado en la concepción de este artículo, en la recopilación, lectura y análisis de la bibliografía citada, en el desarrollo teórico de la investigación y en la redacción. María del Mar Palomares Marín ha recogido los datos del estudio. Francisco Moreno Fernández ha participado en la revisión, en la fundamentación conceptual sobre variación lingüística y en la redacción crítica del artículo.

Financiamiento

Para la realización de este artículo, el investigador Javier Muñoz-Basols ha recibido financiación como Investigador Distinguido Sénior Beatriz Galindo (Ref. BG22/00099), ayuda financiada por el Ministerio de Universidades del Gobierno de España y la Universidad de Sevilla y de los proyectos I+D+i PID2021-123763NA-I00: "Hacia una diacronía de la oralidad/escrituralidad: variación concepcional, traducción y tradicionalidad discursiva en el español y otras lenguas románicas" (DiacOralEs), financiado por MCIN/AEI/10.13039/501100011033/FEDER, UE; DEFINERS: Digital Language Learning of Junior Language Teachers (TED2021-129984A-I00) financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR; "OralGrab. Grabar vídeos y audios para enseñar y aprender" (PID2022-141511NB-100), financiado por la Agencia Estatal de Investigación y por el Ministerio de Ciencia e Innovación del Gobierno de España.

Conflicto de intereses

Los autores no presentan conflicto de interés.

Correspondencia: javier.munoz-basols@mod-langs.ox.ac.uk

Trayectoria académica de los autores

Javier Muñoz-Basols es Investigador Distinguido Sénior Beatriz Galindo en la Universidad de Sevilla (España) y Honorary Faculty Research Fellow en la Universidad de Oxford (Reino Unido). Es investigador principal del "Portal de lingüística hispánica" y coinvestigador principal del proyecto de Humanidades Digitales COMUN-ES (www.comun-es.com). Es cofundador y editor jefe *del Journal of Spanish Language Teaching*, corresponsal del Observatorio Permanente del Hispanismo la Fundación Duques de Soria, miembro del Patronato del Instituto Cervantes, académico correspondiente de la Academia Norteamericana de la Lengua Española (ANLE) y presidente de la Asociación para la Enseñanza del Español como Lengua Extranjera (ASELE).

María del Mar Palomares Marín es doctora en Lingüística Aplicada por la Universidad de Murcia (España). Ha trabajado como Assistant Lecturer in Information Technology en la Technological University Dublin, donde también ha trabajado en el área de español de negocios y español para extranjeros. Cuenta con amplia experiencia en la educación ELE de adultos (programa Lifelong Learning de University College Dublin), y como tutora y teaching fellow de español y literatura para estudiantes universitarios (University College Dublin). Actualmente, es Assistant Professor in Spanish en la Universidad de Limerick, donde participa en varios proyectos de Inteligencia Artificial. Sus líneas de investigación se centran en la tecnología educativa y la Inteligencia Artificial aplicada a la enseñanza de lenguas.

Francisco Moreno Fernández es director del Observatorio Global del Español del Instituto Cervantes, catedrático Alexander von Humboldt en la Universidad de Heidelberg y profesor honorario de la Universidad de Alcalá. Su investigación se ha ocupado de temas relativos a la dialectología y la sociolingüística del español, así como de las lenguas internacionales y la globalización lingüística. Editor jefe de la revista *Spanish in Context* y coeditor de *Journal of Linguistic Geography*. Académico de número de la Academia Europaea y de la Academia Norteamericana de la Lengua Española (ANLE), así como correspondiente de las Academias Cubana, Chilena y Mexicana de la Lengua, y de la Real Academia Española.