INFORMES DEL OBSERVATORIO GLOBAL DEL ESPAÑOL

005

Panorama del español en la inteligencia artificial

Verónica Verdes-Montenegro Díaz

Observatorio Global del Español
Instituto Cervantes

Observatorio Global del Español
Instituto Cervantes

Comité editorial
Luis García Montero (Instituto Cervantes)
Mónica Colomer de Selva (Ministerio de Asuntos Exteriores)
Arturo Colina Aguirre (Fundación para la Transformación de La Rioja)
Mar Pérez Morillo (Ministerio de Transformación Digital)
Francisco Moreno Fernández (Observatorio Global del Español)

Comité Científico
Antonio Briz Gómez (Universidad de Valencia)
Ana M. Cestero Mancera (Universidad de Alcalá)
Álvaro García Santa-Cecilia (Instituto Cervantes)
Elea Giménez Toledo (CSIC)
Asunción Gómez-Pérez (Universidad Politécnica de Madrid)
Ana M. González Mafud (Universidad de La Habana)
Pedro Martín Butragueño (El Colegio de México)
Javier Muñoz-Basols (Universidad de Oxford)
Juan Gabriel Vásquez (Universidad de París-La Sorbona)

Director del Observatorio Global del Español
Francisco Moreno Fernández

Filiación de la autora de esta publicación
Verónica Verdes-Montenegro Díaz (Observatorio Global del Español)

Corrección
María Dolores Pérez Vidal

Maquetación
Diego Sastre Yenes

Coordinación: Francisco Moreno Fernández
Edita: Observatorio Global del Español

Verdes-Montenegro Díaz, V. (2026). *Panorama del español en la inteligencia artificial*. Informes del Observatorio Global del Español, 005. Instituto Cervantes. https://observatoriodelespanol.cevantes.org/documentos/Panorama-del-espanol-en-la-inteligencia-artificial-1.pdf

Fecha de recepción: 21-07-2026
Fecha de aceptación: 03-08-2026

Informes del Observatorio Global del Español
IOGE-005/2026

NIPO (PDF): 110-26-002-0

INFORMES DEL OBSERVATORIO GLOBAL DEL ESPAÑOL

005

Panorama del español en la inteligencia artificial

Verónica Verdes-Montenegro Díaz
Observatorio Global del Español

Observatorio Global del Español
Instituto Cervantes

Contenidos

Resumen 6

Abstract 6

1. Introducción 7

2. Implantación social de la inteligencia artificial en el mundo hispanohablante 10

3. La presencia computacional del español: datos, modelos y repositorios abiertos 18

Common Crawl 19

Hugging Face 20

High Performance Language Technologies 22

4. De los corpus a los modelos: infraestructuras y soberanía tecnológica del español 26

5. Nuevos desafíos y líneas de actuación para el español en la inteligencia artificial 45

6. Conclusiones 53



Resumen

LA INTELIGENCIA ARTIFICIAL se ha convertido en un espacio estratégico para la lengua española. Este informe analiza su posición a partir de cuatro dimensiones: la penetración entre los ciudadanos hispanohablantes, su presencia en iniciativas de datos y modelos abiertos, la disponibilidad de corpus e infraestructuras propias y los desafíos de la nueva etapa digital. Los resultados muestran una demanda social creciente, una base textual considerable y un ecosistema de modelos e iniciativas públicas en expansión. El español presenta fortalezas en traducción, generación textual y tecnologías de la voz, junto con amplias oportunidades de desarrollo en oralidad, imagen, vídeo y evaluación panhispánica. Su principal activo reside en combinar escala demográfica y diversidad lingüística. Transformar este potencial en capacidad tecnológica exige articular los recursos existentes, reforzar la cooperación entre España e Hispanoamérica y consolidar una infraestructura estable, abierta y representativa de la pluralidad del idioma.

Abstract

ARTIFICIAL INTELLIGENCE has become a strategic domain for the Spanish language. This report analyses the position of Spanish across four dimensions: uptake among Spanish-speaking populations, its presence in open-data and open-model initiatives, the availability of dedicated corpora and infrastructure, and the challenges posed by the new digital era. The findings reveal growing social demand, a substantial textual resource base, and an expanding ecosystem of models and public initiatives. Spanish demonstrates particular strengths in translation, text generation, and speech technologies, alongside considerable opportunities for further development in spoken-language applications, image, video, and pan-Hispanic evaluation. Its main asset lies in the combination of demographic scale and linguistic diversity. Turning this potential into technological capacity will require coordinating existing resources, strengthening cooperation between Spain and Spanish America, and consolidating a stable, open infrastructure that represents the full diversity of the language.

1. Introducción

La inteligencia artificial se ha convertido en un espacio de uso y transformación de las lenguas. Los grandes modelos intervienen en actividades cotidianas como buscar información, redactar, traducir, estudiar o resolver tareas profesionales. En este contexto, la posición de un idioma depende de su comunidad de hablantes, pero también de su presencia en los datos, los modelos y las infraestructuras que hacen posibles estas tecnologías.

El español afronta esta transformación desde una posición favorable. Su extensión internacional, su consolidada presencia digital y una comunidad de más de 600 millones de hablantes generan una amplia demanda de herramientas capaces de comprender y producir contenidos en esta lengua. A ello se suma una larga tradición de investigación lingüística, corpus de referencia y un ecosistema creciente de instituciones, proyectos y modelos propios.

La fluidez con la que los sistemas actuales se comunican en español constituye un avance importante y, al mismo tiempo, invita a examinar qué variedades, registros y ámbitos especializados están mejor representados en su uso social. La calidad lingüística de la inteligencia artificial depende tanto del volumen de los datos como de su procedencia, diversidad y documentación.

Este informe analiza la posición del español desde cuatro perspectivas complementarias. En primer lugar, estudia el nivel de adopción de la inteligencia artificial en España y en los países hispanohablantes. A continuación, examina la presencia de la lengua en la tecnología mediante fuentes de datos abiertas. La tercera sección aborda los corpus, las iniciativas institucionales y la evolución de los modelos en español como componentes de una misma infraestructura tecnológica. Finalmente, se identifican los principales desafíos: la representación panhispánica, la oralidad y la multimodalidad, la evaluación desde la propia lengua, la transparencia de los datos y la continuidad de las políticas públicas.

El español cuenta con una base sólida para ampliar su participación en la inteligencia artificial. El propósito general de este informe es analizar cómo sus activos demográficos, lingüísticos e institucionales comienzan a transformarse en capacidad tecnológica y qué líneas de actuación pueden reforzar su presencia como lengua de comunicación, conocimiento e innovación.

Figura 1. El valor demográfico del español y su papel en la inteligencia artificial
Fuente: elaboración propia





2. Implantación social de la inteligencia artificial en el mundo hispanohablante

La expansión reciente de la inteligencia artificial no se explica únicamente por el desarrollo de modelos cada vez más avanzados, sino también por la rapidez con la que estas tecnologías se incorporan a las prácticas cotidianas. Su uso para buscar información, redactar textos, traducir contenidos, estudiar o resolver tareas profesionales convierte a los ciudadanos en uno de los principales agentes de difusión de la IA. En el ámbito hispanohablante, medir el grado de adopción permite observar hasta qué punto estas herramientas han penetrado en sociedades con condiciones económicas, educativas y digitales muy diversas. Los datos ofrecen, por tanto, una primera aproximación a la demanda social de inteligencia artificial en español y a las desigualdades que afectan a su expansión. Conviene recordar, sin embargo, que una utilización elevada no equivale necesariamente a disponer de modelos propios, infraestructuras tecnológicas o recursos lingüísticos suficientes. Implantación y autonomía constituyen dimensiones relacionadas, pero no coincidentes, de la posición que ocupa el español en el nuevo ecosistema de la inteligencia artificial.

Según los datos publicados por el Microsoft AI Economy Institute, los resultados correspondientes al primer trimestre de 2026 muestran un ecosistema hispanohablante extenso y articulado a diferentes velocidades.

Las tasas de adopción oscilan entre el 44,2 % registrado en España y el 6,7 % de Cuba, lo que refleja una presencia ya extendida de la inteligencia artificial en buena parte de los países analizados. Tras España se sitúan Costa Rica, con una tasa del 28,5 %; Uruguay, con el 24,6 %; Colombia, con el 24,5 %; Panamá, con el 23,3 %; Chile, con el 22,7 %; Argentina, con el 21,9 %, y México, con el 20,1 %. La presencia de siete países latinoamericanos por encima del 20 % revela la formación de varios polos de implantación distribuidos por distintos territorios de la región. En el siguiente tramo se encuentran Ecuador, con una tasa del 19,5 %; El Salvador, con el 18,3 %; Guatemala y Perú, con el 16,4 %; Honduras, con el 14 %; Bolivia, con el 12,7 %; Paraguay, con el 12,2 %; Nicaragua, con el 11,8 %; Venezuela, con el 10,3 %, y Guinea Ecuatorial, con el 10,1 %. Cuba, con el 6,7 %, completa la relación de países incluidos en el indicador. Esta variabilidad en las cifras permite reconocer distintos ritmos de penetración de la IA y muestra la influencia que ejercen la conectividad, las competencias digitales, la estructura productiva y la disponibilidad de servicios tecnológicos en su difusión.

Figura 2. Representación global de la adopción de la IA en países hispanohablantes
Fuente: elaboración propia a partir de *Global AI Adoption in 2025–2026*. AI Economy Institute. Disponible en: https://www.microsoft.com/en-us/corporate-responsibility/topics/ai-economy-institute/reports/global-ai-adoption-2026-q1/ (consultado el 22 de junio de 2026)

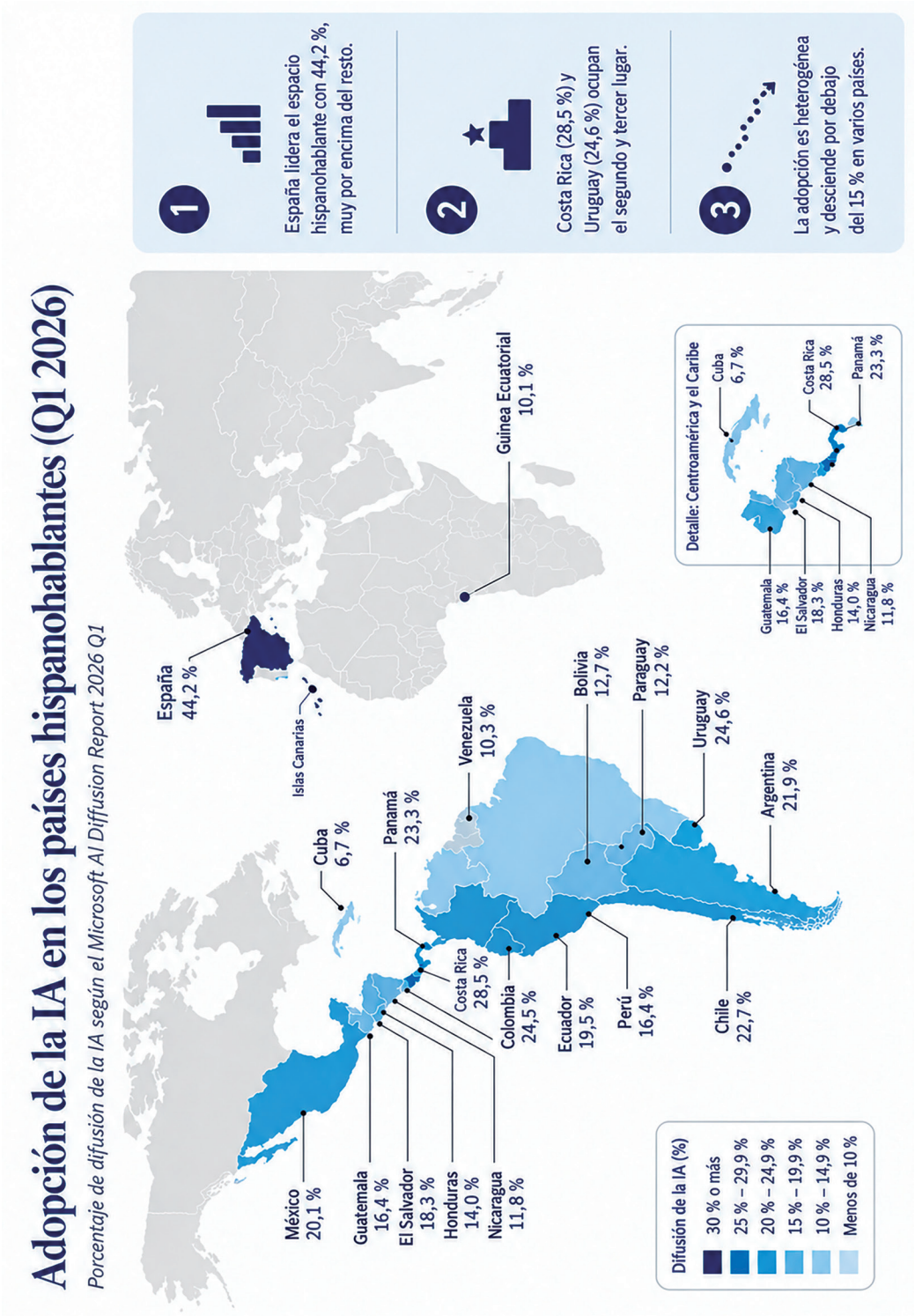
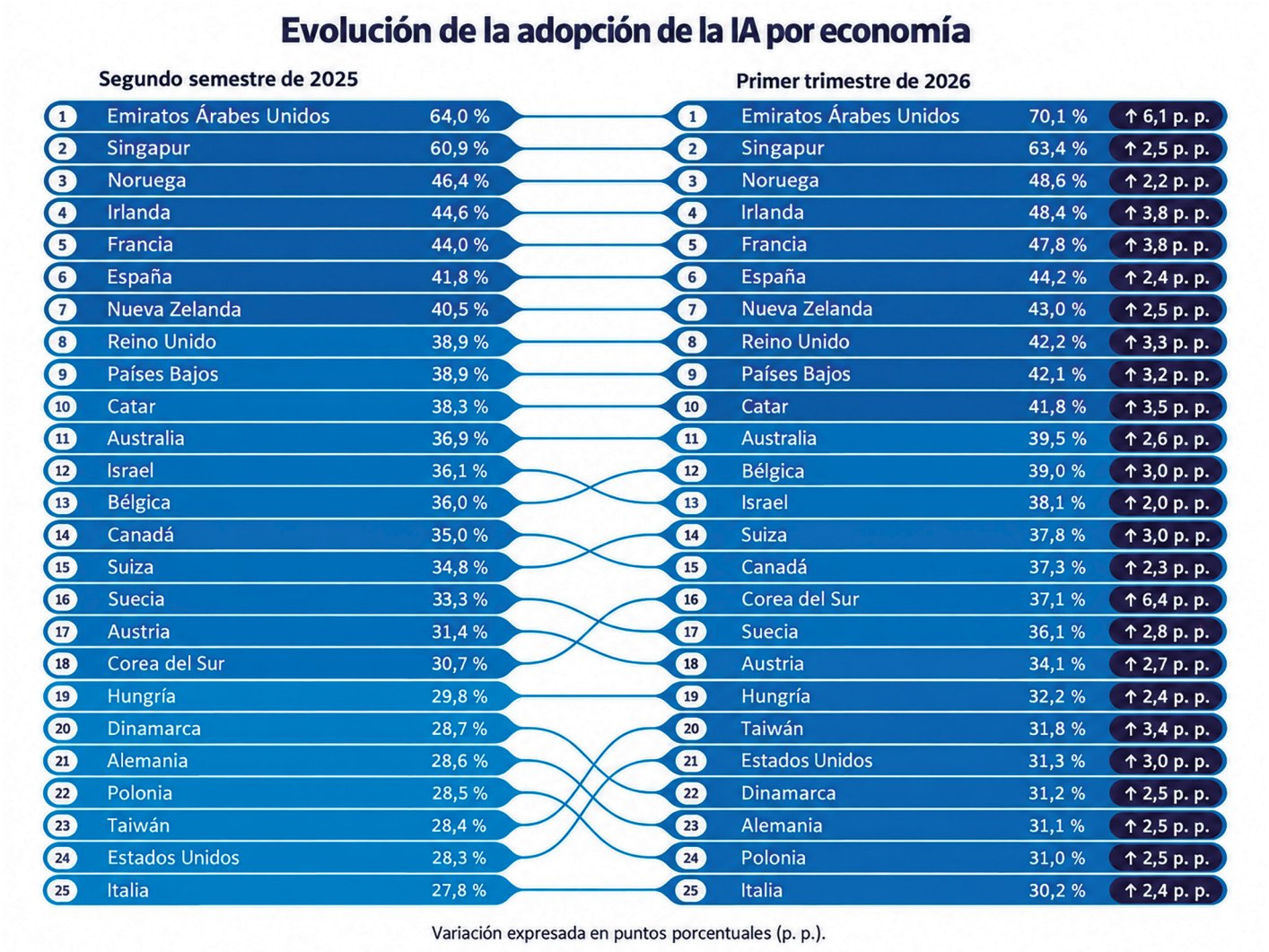


Tabla 1. Tasa de implantación de la inteligencia artificial generativa en los países hispanohablantes (primer trimestre de 2026)
Fuente: elaboración propia a partir de datos de Microsoft AI Economy Institute (2026), Global AI Diffusion: Q1 2026 Trends and Insights

País	Tasa de adopción
España	44,2 %
Costa Rica	28,5 %
Uruguay	24,6 %
Colombia	24,5 %
Panamá	23,3 %
Chile	22,7 %
Argentina	21,9 %
México	20,1 %
Ecuador	19,5 %
El Salvador	18,3 %
Guatemala	16,4 %
Perú	16,4 %
Honduras	14,0 %
Bolivia	12,7 %
Paraguay	12,2 %
Nicaragua	11,8 %
Venezuela	10,3 %
Guinea Ecuatorial	10,1 %
Cuba	6,7 %

Es importante destacar que el indicador elaborado por Microsoft mide la proporción de personas de entre 15 y 64 años que han utilizado un producto de inteligencia artificial generativa durante el periodo analizado. Las estimaciones se obtienen a partir de datos de uso agregados y anonimizados, ajustados según factores como la penetración de internet, la disponibilidad de dispositivos, la cuota de mercado de los sistemas operativos y el tamaño de la población.

Figura 3. Ranking de adopción de IA generativa en economías seleccionadas. Comparativa entre el segundo semestre de 2025 y el primer trimestre de 2026
Fuente: Microsoft AI Economy Institute (2026). Global AI Adoption Report Q1 2026. Disponible en: https://www.microsoft.com/en-us/corporate-responsibility/topics/ai-economy-institute/reports/global-ai-adoption-2026-q1/



No se trata, por tanto, de una encuesta de percepción, sino de una aproximación al uso efectivo de estas herramientas, aunque su dependencia de la telemetría de Microsoft aconseja interpretar los resultados dentro de los límites de esta metodología.

En cuanto a la difusión de la IA en las distintas economías mundiales, Microsoft sitúa a España como el único país hispanohablante entre los primeros puestos de su clasificación de adopción. Con una tasa del 44,2 %, España supera ampliamente el promedio mundial, situado en el 17,8 %, y ocupa la sexta posición internacional, por detrás de los Emiratos Árabes Unidos, Singapur, Noruega, Irlanda y Francia. Su evolución reciente confirma, además, la continuidad de este proceso: entre el segundo semestre de 2025 y el primer trimestre de 2026, la proporción de usuarios aumentó del 41,8 % al 44,2 %, lo que representa un incremento de 2,4 puntos porcentuales.

La comparación internacional permite precisar el alcance de estos datos. España ocupa la sexta posición mundial, por detrás de los Emiratos Árabes Unidos (70,1 %), Singapur (63,4 %), Noruega (48,6 %), Irlanda (48,4 %) y Francia (47,8 %). Se integra así en un grupo reducido de economías en las que el uso de la IA generativa se aproxima o supera el 50 % de la población en edad laboral. Su presencia en este tramo de la clasificación resulta especialmente significativa, puesto que la difusión mundial de estas tecnologías continúa siendo muy desigual y se concentra, sobre todo, en economías con elevados niveles de conectividad, capacitación digital y renta.

La posición de España no responde únicamente a un valor elevado en un momento determinado. Entre el segundo semestre de 2025 y el primer trimestre de 2026, la proporción de usuarios pasó del 41,8 % al 44,2 %, lo que representa un incremento de 2,4 puntos porcentuales. Aunque el aumento fue inferior al registrado por Irlanda o Francia —3,8 puntos en ambos casos—, resulta relevante porque se produjo sobre una base de adopción ya considerable. España mantuvo, además, la sexta posición en los dos periodos analizados. Esta continuidad indica que la inteligencia artificial generativa se está integrando de manera sostenida en los usos sociales, educativos y profesionales de la población, y no como consecuencia de un crecimiento meramente coyuntural.

Una perspectiva complementaria sobre la implantación social de la IA en los usuarios hispanohablantes es la que nos ofrece el Índice Latinoamericano de Inteligencia Artificial 2025 (Soto *et al.*, 2025, pp. 136-139). Su indicador sobre el uso de la IA generativa en la web, elaborado a partir del tráfico registrado en abril de 2025, sitúa a Chile, Costa Rica, Perú, Uruguay, Panamá, Ecuador, Argentina y Colombia en las primeras posiciones entre los países hispanohablantes de América Latina. Aunque esta metodología difiere de la empleada por Microsoft —que contabiliza visitas a páginas de IA generativa en relación con el número de usuarios de internet en lugar de estimar usuarios únicos—, sus resultados confirman la amplia incorporación de estas herramientas a las prácticas digitales de la población.

Figura 4. Representación de la difusión de la inteligencia artificial en las economías líderes (primer semestre de 2026)
Fuente: Microsoft (2026). Global AI Adoption in 2025–2026. AI Economy Institute. Disponible en: https://www.microsoft.com/en-us/corporate-responsibility/topics/ai-economy-institute/reports/global-ai-adoption-2026-q1/ (consultado el 22 de junio de 2026)

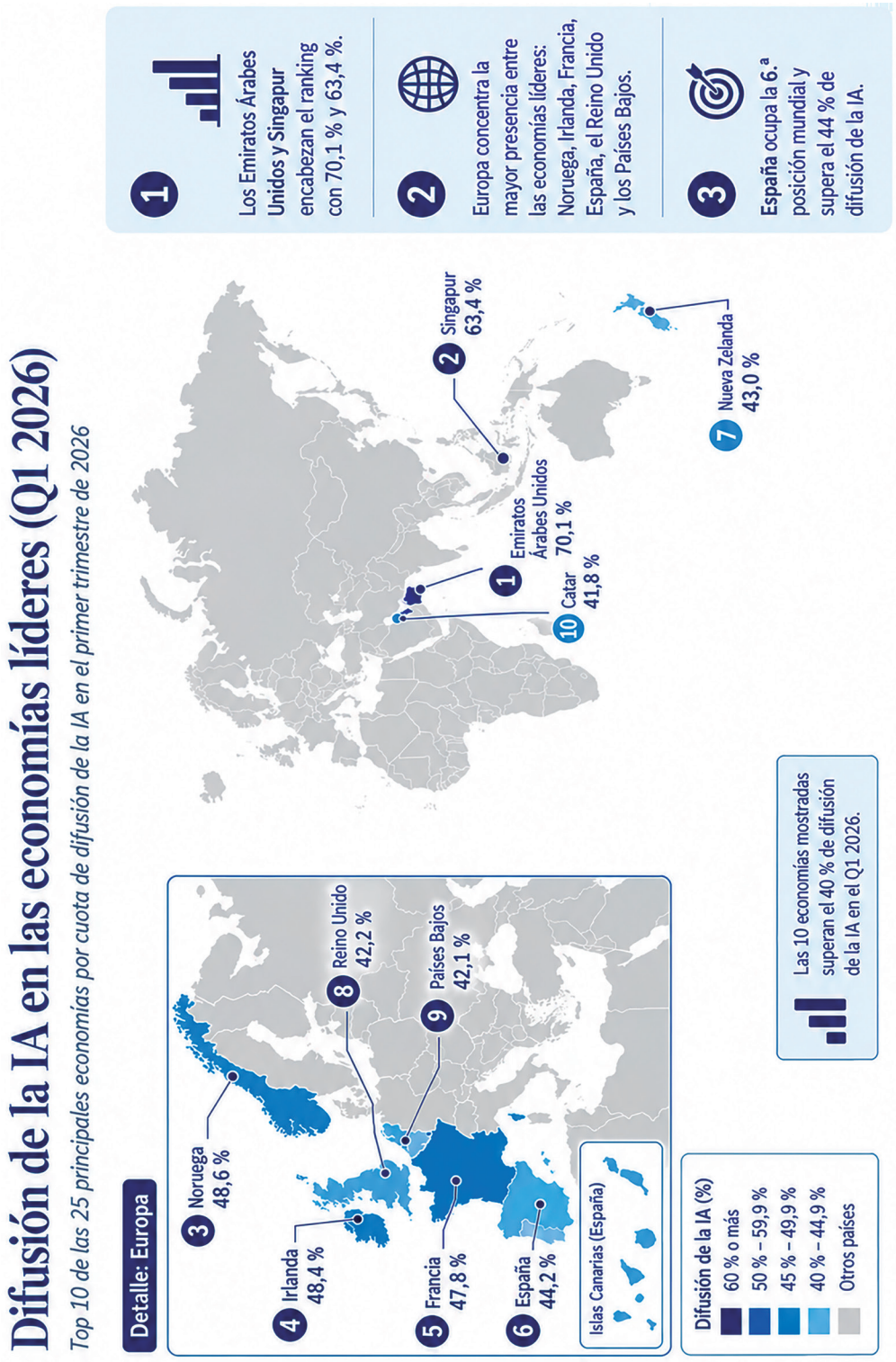
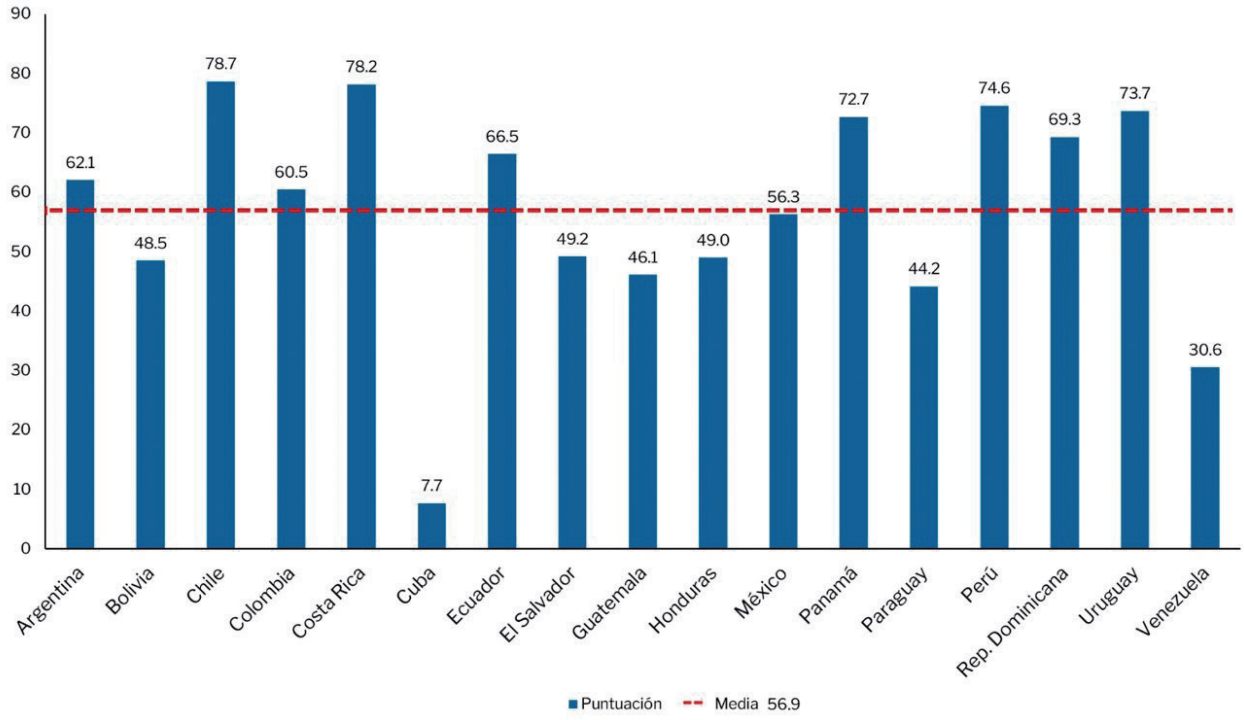


Gráfico 1. Frecuencia de uso de la IA generativa en los países hispanohablantes de América Latina (2025)
Fuente: elaboración propia a partir de datos de Soto *et al.* (2025), Índice Latinoamericano de Inteligencia Artificial 2025, pp. 136-137



Estas cifras no pueden compararse directamente con las tasas de Microsoft, ya que no corresponden a porcentajes de población usuaria ni a recuentos de usuarios únicos, sino que miden el tráfico web dirigido a servicios de IA generativa en relación con el número de usuarios de internet de cada país. No obstante, el gráfico permite confirmar la amplia incorporación de estas herramientas a las prácticas digitales de los países hispanohablantes analizados. Con una media de 56,94 puntos, el conjunto presenta una demanda social considerable, aunque con diferencias apreciables entre países en cuanto a la intensidad efectiva de uso de estas tecnologías.

La lectura lingüística de estos datos es especialmente importante. Cuanto mayor sea el número de hablantes que utilizan herramientas de IA en español, mayor será también la demanda de sistemas capaces de comprender sus variedades, registros, acentos, necesidades comunicativas y contextos culturales. Sin embargo, las diferencias en los niveles de adopción muestran que la presencia del español en la IA no puede medirse únicamente en términos globales, sino que debe atender también a la distribución territorial de los usuarios y a las diversas condiciones en las que acceden a estas tecnologías.



3. La presencia computacional del español: datos, modelos y repositorios abiertos

Los datos de adopción examinados en el apartado anterior dibujan un espacio hispanohablante en el que la inteligencia artificial ya forma parte de las prácticas digitales de millones de ciudadanos, aunque con intensidades muy distintas según los países. Esa expansión permite medir la demanda social de estas tecnologías, pero deja abierta una cuestión más profunda: ¿qué lugar ocupa realmente el español en los sistemas que las hacen posibles? Una lengua puede contar con una comunidad extensa de usuarios y, sin embargo, depender de modelos concebidos, entrenados y evaluados principalmente en otras lenguas. Pasar de la adopción al liderazgo computacional exige, por tanto, desplazar la mirada desde quienes utilizan la inteligencia artificial hacia los datos, los modelos y las infraestructuras que determinan cómo esta comprende y genera el idioma español.

Este desplazamiento plantea una dificultad metodológica. No existe en la actualidad una base pública, homogénea y comparable que permita conocer qué proporción de español contiene cada uno de los grandes modelos comerciales. Los principales desarrolladores difunden información limitada sobre la composición lingüística de sus corpus de entrenamiento y sobre los procesos de selección, limpieza y ponderación de los datos. En consecuencia, tampoco es posible determinar con precisión qué variedades geográficas, registros sociales o ámbitos especializados están representados. Ante esta falta de datos, el análisis debe recurrir a indicadores indirectos: grandes colecciones abiertas de datos web y plataformas de modelos que, sin reproducir la arquitectura interna de los sistemas propietarios, permiten aproximarse a las condiciones materiales desde las que el español participa en el ecosistema de la IA.

Por ello, el análisis de la presencia del español puede partir del estudio de los recursos disponibles en abierto, que también se emplean para entrenar los grandes modelos de inteligencia artificial.

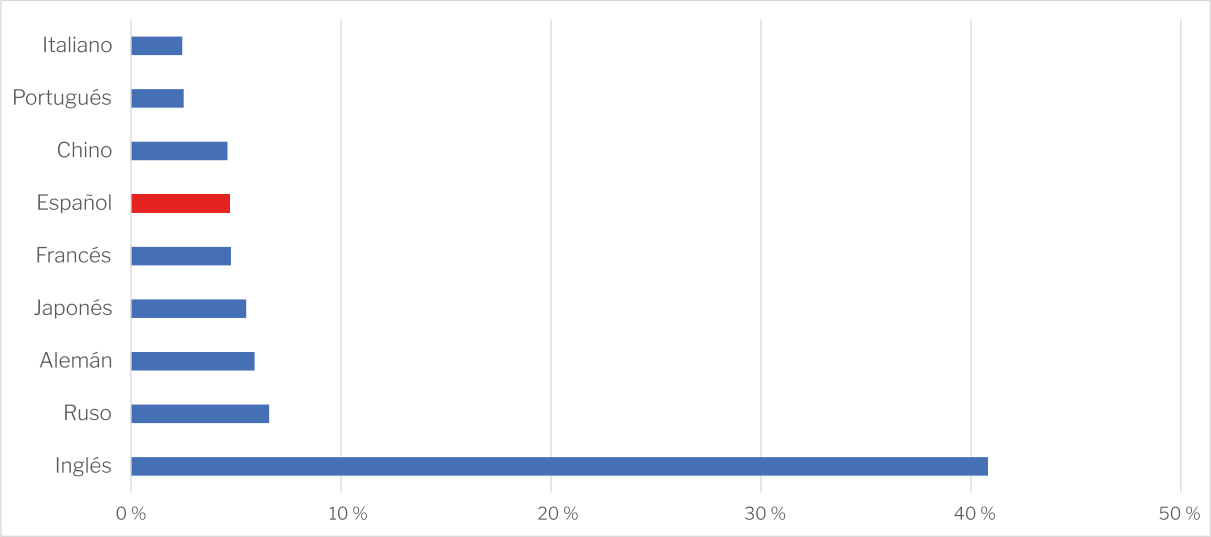


Common Crawl

El primero de estos indicadores es Common Crawl, una de las mayores infraestructuras abiertas de datos web. Su interés no reside en que revele el corpus concreto de ningún modelo comercial, sino en que permite observar la disponibilidad relativa de contenidos en las lenguas que alimentan una parte sustancial del desarrollo contemporáneo de la inteligencia artificial. Se trata, por ello, de un indicador indirecto: las empresas pueden seleccionar otras fuentes y aplicar filtros, deduplicaciones[1] o ponderaciones diferentes. Aun con esta cautela, la distribución lingüística de Common Crawl ofrece una referencia útil para situar al español en la economía global de los datos digitales.

En la última medición disponible de Common Crawl, correspondiente a CC-MAIN-2026-25, el español representa el 4,68 % de las páginas HTML identificadas por lengua principal. Esta proporción sitúa al español entre las lenguas con mayor presencia en el rastreo web global, en un nivel muy próximo al francés (4,77 %) y por encima del chino (4,58 %) y del portugués (2,51 %). Sin embargo, la distancia con el inglés sigue siendo estructural: el inglés concentra el 40,82 % de los documentos, es decir, casi nueve veces más que el español.

Gráfico 2. Distribución de lenguas en Common Crawl, CC-MAIN-2026-25
Fuente: elaboración propia a partir de datos de Common Crawl



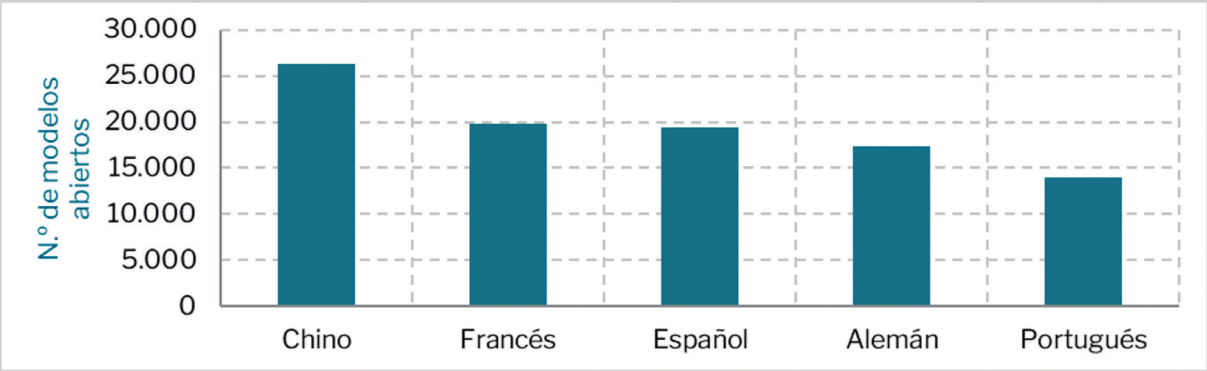
1 Proceso de depuración que consiste en reducir los datos redundantes al eliminar copias adicionales de la misma información. El objetivo de la deduplicación de datos es reducir las necesidades continuas de almacenamiento de las organizaciones» (Powell y Smalley, 2025; véase https://www.ibm.com/es-es/think/topics/data-deduplication).

Hugging Face

Si Common Crawl permite aproximarse a la disponibilidad de datos, Hugging Face desplaza la observación hacia los modelos y las herramientas que circulan en el entorno abierto. La plataforma reúne modelos, conjuntos de datos, aplicaciones para realizar pruebas y recursos reutilizables por universidades, empresas, administraciones y comunidades de desarrolladores. Aunque su catálogo no abarca toda la inteligencia artificial existente ni equivale al mercado de los grandes sistemas comerciales, permite analizar hasta qué punto una lengua participa en los circuitos abiertos de creación, evaluación y transferencia tecnológica.

Los datos recopilados hasta junio de 2026 revelan que, si realizamos una comparación en las principales lenguas distintas del inglés, la posición del español resulta relativamente sólida.

Gráfico 3. Modelos disponibles en Hugging Face etiquetados en las principales lenguas distintas del inglés
Fuente: elaboración propia a partir de datos recogidos en Hugging Face hasta junio de 2026



El español cuenta con 19.360 modelos abiertos etiquetados en esta lengua, una cifra muy próxima a los 19.832 del francés y solo inferior a los 26.309 del chino. A su vez, supera al alemán, con 17.314 modelos, y al portugués, con 14.045. El gráfico muestra, por tanto, que el español forma parte del grupo de lenguas con mayor presencia en el ecosistema abierto de Hugging Face. Esta posición comparativa debe interpretarse, no obstante, teniendo en cuenta que los modelos etiquetados en español representan una proporción reducida del catálogo total de la plataforma.

Por otra parte, el análisis de la presencia del español en las distintas tareas para las que se desarrollan estos modelos aporta una perspectiva más precisa y permite matizar la imagen general ofrecida por los datos agregados.

Tabla 2. Indicadores de modelos disponibles en Hugging Face etiquetados por lengua española
Fuente: elaboración propia a partir de datos recogidos en Hugging Face hasta junio de 2026

Indicador Hugging Face	Modelos en español	Peso sobre el total	Posición entre lenguas comparadas
Total de modelos abiertos etiquetados	19.360	0,66 %	4.ª
Generación de texto	7.261	1,93 %	4.ª
Traducción	1.607	10,82 %	3.ª
Clasificación de textos	757	0,64 %	2.ª
Respuestas a preguntas	110	0,79 %	4.ª
Similitud de oraciones	431	0,11 %	4.ª
Texto a voz	677	11,04 %	3.ª
Reconocimiento automático del habla	1.087	3,33 %	4.ª
Texto a imagen	25	0,02 %	5.ª
Texto a vídeo	0	0,00 %	6.ª

El español muestra una implantación sólida en ámbitos consolidados del procesamiento del lenguaje natural, como la traducción automática, la clasificación de textos, la generación textual y las tecnologías del habla. En el ámbito de la traducción, los modelos abiertos etiquetados en español representan el 10,82 % del total, una cuota muy superior al 0,66 % que corresponde a esta lengua en el conjunto de la plataforma. También destaca su presencia en el reconocimiento automático del habla, con 1.087 modelos, y en la conversión de texto a voz, con 677. Esta especialización constituye un activo relevante para desarrollar asistentes conversacionales, herramientas educativas, sistemas de accesibilidad y aplicaciones que requieren una interacción oral fluida.

La situación es distinta en las modalidades generativas más recientes. Hugging Face registra únicamente 25 modelos de texto a imagen etiquetados en español y ninguno de texto a vídeo en los datos analizados. Esta diferencia resulta significativa, pues, a medida que la inteligencia artificial integra texto, voz, imagen y vídeo en una misma arquitectura, la posición de las lenguas dependerá también de su capacidad para incorporarse a

estos entornos multimodales. El español cuenta con una presencia consolidada en las tareas textuales y orales, pero todavía ocupa un espacio reducido en los desarrollos que anticipan la siguiente fase de la IA generativa.

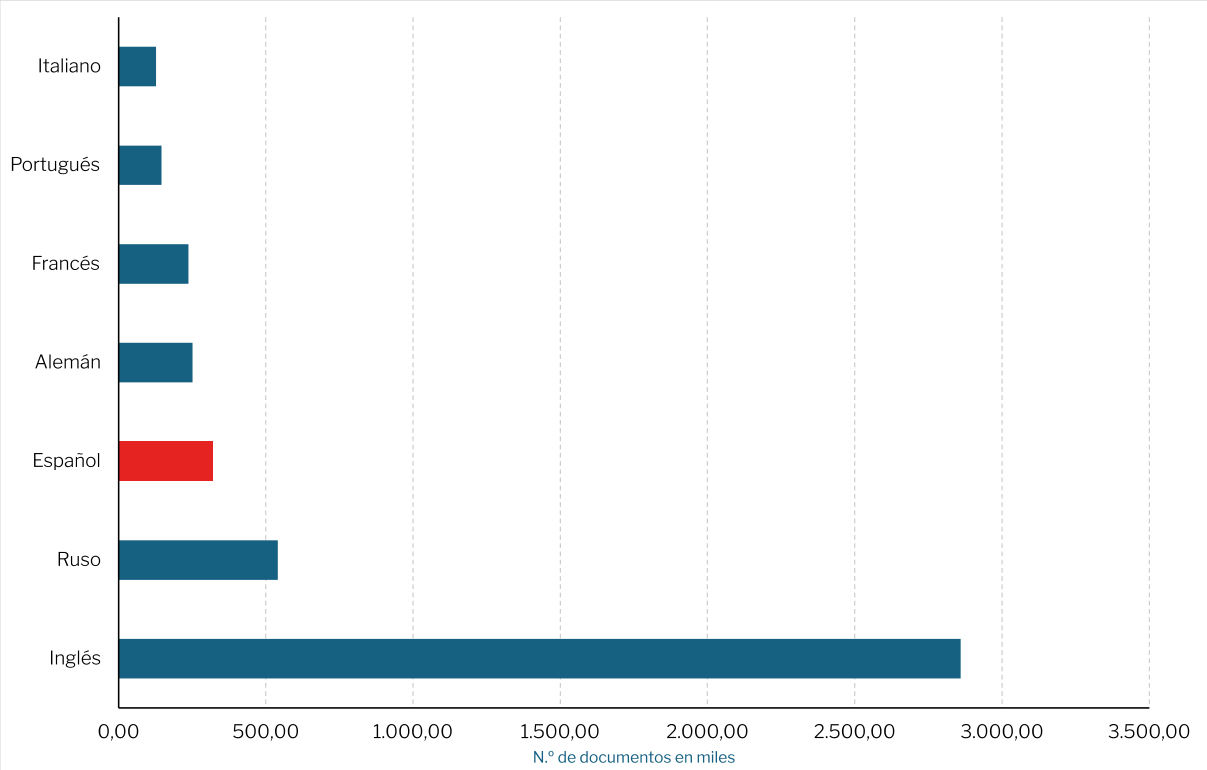
High Performance Language Technologies

La tercera referencia, High Performance Language Technologies (HPLT), permite pasar del volumen bruto de datos disponibles en la web a su procesamiento para el desarrollo de tecnologías lingüísticas. A diferencia de Common Crawl, HPLT aplica procedimientos específicos de extracción, identificación de lenguas, deduplicación y limpieza. Su versión 2.0 reúne corpus monolingües en 193 lenguas y ofrece dos colecciones: una deduplicada, de 21 TB, y otra limpia, de 15 TB. El propio proyecto recomienda esta última cuando no se prevén procesos adicionales de depuración. Este indicador resulta especialmente pertinente porque muestra tanto la cantidad de contenido en español que circula por internet como la proporción que reúne las condiciones necesarias para convertirse en un recurso potencial de entrenamiento y evaluación.

Dentro de esta infraestructura, el español ocupa una posición destacada. La colección deduplicada contiene 838,93 millones de documentos, 414,23 mil millones de palabras, 2,53 billones de caracteres y 22,17 mil millones de segmentos. Tras el proceso de limpieza se conservan 503,07 millones de documentos, 321,95 mil millones de palabras, 1,95 billones de caracteres y 12,12 mil millones de segmentos. Por número de palabras la colección limpia en español supera a las de francés, alemán, portugués e italiano, aunque se mantiene a una distancia considerable del inglés y se sitúa también por debajo del ruso en esta plataforma.



Gráfico 4. Presencia comparada del español en HPLT v2.0.
Fuente: elaboración propia a partir de datos de HPLT Datasets v2.0.



La comparación confirma que el español dispone de una masa textual de gran escala y potencialmente utilizable para entrenar modelos. Ahora bien, el paso de la colección deduplicada a la limpia reduce el volumen inicial y recuerda que la abundancia de contenidos no equivale por sí sola a calidad computacional. Aproximadamente tres cuartas partes de las palabras se conservan después del filtrado; el resto queda fuera por problemas de duplicación, ruido u otros criterios de depuración.

Este resultado incorpora una dimensión cualitativa que las cifras agregadas no permiten valorar por sí solas. En una lengua internacional y pluricéntrica como el español, además de reunir cientos de miles de millones de palabras, resulta esencial conocer su procedencia, garantizar su trazabilidad y comprobar que las distintas variedades territoriales, los registros sociales y los ámbitos especializados estén adecuadamente representados. Un corpus voluminoso puede reproducir los desequilibrios presentes en la web y ofrecer una imagen parcial del idioma cuando sus datos se concentran en unos pocos países, medios de comunicación o tipos de texto.

Considerados en conjunto, los tres indicadores permiten trazar un panorama común sobre la presencia del español en la IA y señalan ámbitos con un

amplio margen de desarrollo. Common Crawl muestra que el español cuenta con una presencia relevante en la web; HPLT confirma que una parte considerable de esos contenidos puede transformarse en datos procesados a gran escala; y Hugging Face refleja la existencia de una comunidad capaz de convertir estos recursos en modelos y aplicaciones abiertas.



4. De los corpus a los modelos: infraestructuras y soberanía tecnológica del español

Los indicadores analizados en el apartado anterior confirman que el español cuenta con una presencia significativa en las grandes colecciones de datos web y en los repositorios abiertos de inteligencia artificial. Sin embargo, la disponibilidad de contenidos y modelos no permite determinar por sí sola qué variedades del español aprenden los sistemas ni en qué medida la comunidad hispanohablante puede intervenir en su desarrollo. Para responder a estas cuestiones resulta necesario analizar conjuntamente los corpus, las infraestructuras tecnológicas, las políticas institucionales y los modelos construidos a partir de estos recursos. Todos estos elementos forman parte de una misma cadena: los datos aportan la base lingüística; las instituciones establecen las condiciones necesarias para conservarlos y facilitar su acceso; y los modelos transforman esos recursos en aplicaciones y capacidades tecnológicas.

Desde esta perspectiva, la soberanía tecnológica no debe entenderse como aislamiento ni como autosuficiencia absoluta. Se refiere, más bien, a la capacidad de una comunidad para decidir qué datos utiliza, cómo los documenta, qué variedades lingüísticas representa, con qué criterios entrena y evalúa sus modelos y con qué fines pone estos sistemas a disposición de los ciudadanos, las empresas y las administraciones públicas. Una lengua puede contar con millones de hablantes y grandes cantidades de texto digital y, al mismo tiempo, depender de tecnologías desarrolladas en otros contextos. La diferencia entre presencia y soberanía reside, precisamente, en la capacidad para transformar esos recursos lingüísticos en conocimiento, infraestructura y desarrollo propio.



Los corpus como infraestructura lingüística

Los corpus ya no son únicamente depósitos organizados de textos o muestras de habla. Son espacios en los que los sistemas identifican regularidades gramaticales, establecen relaciones semánticas, aprenden asociaciones entre palabras y contextos y se aproximan a los usos comunicativos de una comunidad. Por tanto, su calidad, diversidad y representatividad condicionan tanto el rendimiento de las aplicaciones como la posición que una lengua puede ocupar en el ecosistema tecnológico.

El *Informe del estado actual de los corpus en español, lenguas cooficiales y variantes del español* (2023), elaborado por el Ministerio de Asuntos Económicos y Transformación Digital y el Instituto Cervantes, identifica más de 190 corpus relevantes desarrollados para el español, las lenguas cooficiales y las variedades del español, por lo que constituye el inventario más completo realizado hasta la fecha. Estos recursos abarcan corpus textuales, orales, multimodales, terminológicos y especializados, destinados tanto a la investigación como al desarrollo industrial de tecnologías del lenguaje.

Tabla 3. Situación general de los corpus en español
Fuente: elaboración propia a partir del Informe del estado actual de los corpus en español, lenguas cooficiales y variantes del español (2023)

Indicador	Situación actual
Corpus identificados oficialmente	Más de 190
Tipologías representadas	Texto, voz, multimodal, terminológico, especializado
Cobertura geográfica	España, Hispanoamérica y variedades hispánicas
Lenguas incluidas	Español, catalán, gallego, euskera y lenguas de signos
Nivel de madurez	Alto en corpus textuales; medio en corpus orales y multimodales

Gracias a una larga tradición de lingüística de corpus y al trabajo de instituciones académicas, científicas y públicas, el español dispone de recursos de referencia que documentan la lengua en sus múltiples manifestaciones locales, sociales y temáticas. Estos corpus proporcionan millones de palabras necesarias para que los sistemas aprendan patrones gramaticales, relaciones semánticas, variedades lingüísticas y conocimientos especializados.

Corpus textuales

Tabla 4. Corpus textuales en español
Fuente: elaboración propia a partir del *Informe del estado actual de los corpus en español, lenguas cooficiales y variantes del español* (2023)

Nombre	Descripción / Institución	Tipo	Cobertura geográfica	Última actualización publicada	Tamaño
Corpus de aprendices de español (CAES)	Instituto Cervantes / Universidad de Santiago de Compostela	Escrito	Aprendices de español	2014	1 millón de palabras
Corpus de Referencia de la Lengua Española en la Argentina	Universidad Autónoma de Madrid	Escrito	Español de Argentina	1992	2 millones de palabras
Corpus de Referencia del Español Actual (CREA)	Real Academia Española	Escrito (90 %) / Oral (10 %)	América (50 %) / España (50 %)	junio de 2008	160 millones de palabras
Corpus de Referencia del Español de Chile	Universidad Autónoma de Madrid	Escrito	Español de Chile	1991-1994	2 millones de palabras

Nombre	Descripción / Institución	Tipo	Cobertura geográfica	Última actualización publicada	Tamaño
Corpus del Diccionario Histórico de la Lengua Española (CDH)	Real Academia Española	Escrito - Histórico	América / España	octubre de 2021	355 millones de palabras
Corpus del Español Contemporáneo de España y de Hispanoamérica (CUMBRE)	Editorial SGEL	Escrito / Oral (transcripciones)	América (35 %) / España (65 %)	1995	20 millones de palabras
Corpus del Español del Siglo XXI (CORPES XXI)	Real Academia Española	Escrito (90 %) / Oral (10 %)	América (70 %) / España (30 %)	junio de 2026	455 millones de palabras
Corpus del Español en los Estados Unidos (CORPEEU)	Academia Norteamericana de la Lengua Española / Instituto Cervantes - Harvard	Escrito (80 %) / Oral (10 %) / Tecleada (10 %)	Español de EE.UU.	N/C	92 millones de palabras
Corpus del Español Mexicano Contemporáneo (CEMC)	Colegio de México	General	Español de México	2018	100 millones de palabras
Corpus Diacrónico del Español (CORDE)	Real Academia Española	Escrito - Histórico	América (26 %) / España (74 %)	N/C	250 millones de palabras
Corpus Diacrónico y Diatópico del Español de América (CORDIAM)	Academia Mexicana de la Lengua / Asociación de Academias	Escrito	Español de América	2015	7,3 millones de palabras
Corpus Dinámico del Castellano de Chile (CODICACH)	Universidad Católica de Chile	Escrito	Español de Chile	N/C	900 millones de palabras

Nombre	Descripción / Institución	Tipo	Cobertura geográfica	Última actualización publicada	Tamaño
Corpus Lingüístico español-gallego de la Universidad de Vigo (CLUVI)	Universidad de Vigo	Textos paralelos	Español, gallego, francés y otros	2008	49 millones de palabras
Corpus MAVIR	Consorcio MAVIR	Español e inglés	—	2008	103.479 palabras

En la amplia nómina de corpus textuales destacan los recursos impulsados por la Real Academia Española. La versión 1.5 del Corpus del Español del Siglo XXI (CORPES XXI), publicada en junio de 2026, reúne casi 420.000 documentos y 455 millones de formas ortográficas procedentes de textos escritos y transcripciones orales, diez millones más que la entrega anterior. Junto con el Corpus de Referencia del Español Actual (CREA), que contiene aproximadamente 160 millones de formas, constituye un recurso fundamental para el estudio del español contemporáneo y su diversidad geográfica y discursiva.

A estos recursos se suman los grandes corpus históricos, como el Corpus Diacrónico del Español (CORDE) y el Corpus del Diccionario Histórico de la Lengua Española (CDH), que permiten incorporar la dimensión temporal al análisis computacional del idioma. Aunque estos corpus fueron concebidos con fines filológicos, su riqueza resulta cada vez más valiosa para aplicaciones relacionadas con la evolución semántica, la detección de cambios lingüísticos o el desarrollo de sistemas de procesamiento histórico de textos.

En el ámbito de los corpus masivos procedentes de la web destacan el Corpus del Español desarrollado por la Universidad Brigham Young y la Illinois State University, así como el Spanish Web Corpus (esTenTen) de Sketch Engine. Con miles de millones de palabras recopiladas mediante técnicas de rastreo web, estos recursos ofrecen una escala especialmente adecuada para los requerimientos de los sistemas actuales de inteligencia artificial, cuyos modelos demandan volúmenes de datos considerablemente superiores a los utilizados tradicionalmente en lingüística de corpus.

Junto a los grandes corpus generales, el español cuenta con recursos especializados que aportan diversidad temática y geográfica. Entre ellos figuran el Corpus del Español Mexicano Contemporáneo (CEMC),

el Corpus Dinámico del Castellano de Chile (CODICACH), el Corpus Diacrónico y Diatópico del Español de América (CORDIAM) o el Corpus del Español en los Estados Unidos (CORPEEU). Estos materiales resultan especialmente relevantes para reducir sesgos regionales y favorecer el desarrollo de modelos que representen adecuadamente la pluralidad del espacio hispanohablante.

Corpus orales

La expansión de los asistentes virtuales, los sistemas de reconocimiento automático del habla y las tecnologías conversacionales ha otorgado una relevancia creciente a los corpus orales. Mientras los recursos escritos permiten aprender estructuras gramaticales, vocabulario y relaciones semánticas, los datos de voz incorporan pronunciaciones, acentos, entonaciones y fenómenos propios de la interacción espontánea. Sin esta dimensión la representación computacional del español permanece necesariamente incompleta.

Entre los recursos disponibles destacan el corpus Val.Es.Co., centrado en la conversación coloquial; ESLORA, desarrollado por la Universidad de Santiago de Compostela; el Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC); y PRESEEA, que reúne entrevistas sociolingüísticas procedentes de numerosas ciudades de España y América. Estos corpus documentan fenómenos poco frecuentes en la escritura, como interrupciones, solapamientos, vacilaciones, marcadores discursivos, variaciones prosódicas y estrategias de interacción.

Otros recursos, como el Corpus de Habla Infantil Espontánea del Español (CHIEDE) o el Corpus Oral de las Lenguas Romances (C-ORAL-ROM), amplían la representación de perfiles de hablantes y situaciones comunicativas. Su incorporación al desarrollo tecnológico resulta especialmente relevante para construir aplicaciones capaces de reconocer distintas edades, pronunciaciones, ritmos de habla y condiciones de interacción.

Por el volumen de datos orales reunidos, PRESEEA ocupa una posición singular debido a su amplia cobertura geográfica y a la homogeneidad de su metodología. La recopilación de entrevistas comparables en distintas comunidades urbanas permite observar el habla real de grupos sociales, generaciones y niveles educativos diversos. Este equilibrio convierte al proyecto en un recurso especialmente valioso para entrenar y evaluar sistemas de reconocimiento del habla y modelos conversacionales menos dependientes de unas pocas variedades dominantes.

Desde la perspectiva de la IA, la disponibilidad de corpus orales amplios y representativos constituye una condición indispensable para desarrollar tecnologías de voz que reflejen la diversidad de acentos, variedades dialectales y contextos sociolingüísticos del español.

Tabla 5. Corpus orales en español
Fuente: elaboración propia a partir de datos del *Informe del estado actual de los corpus en español, lenguas cooficiales y variantes del español* (2023)

Nombre	Descripción / Institución	Tipo	Cobertura geográfica	Última actualización publicada	Tamaño
Corpus de Habla Infantil Espontánea del Español (CHIEDE)	Universidad Autónoma de Madrid	Conversaciones y entrevistas	Español	2008	74 horas / 58.163 palabras
Corpus Oral de las Lenguas Romances (C-ORAL-ROM)	Universidad de Florencia	Habla espontánea	Francés, italiano, portugués y español	2004	300.000 palabras por lengua
Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC)	Universidad Autónoma de Madrid	Oral (transcripciones)	Español de España	1992	1,1 millón de palabras
Corpus para el estudio del español oral (ESLORA)	Universidad de Santiago de Compostela	Entrevistas y conversaciones	Español (Galicia)	enero de 2026	80 horas
Valencia Español Coloquial (Val.Es.Co)	Universidad de Valencia	Oral (coloquial)	España	mayo de 2026	66 conversaciones
Proyecto para el Estudio Sociolingüístico del Español de España y de América (PRESEEA)	Universidad de Alcalá	Entrevistas	España e Iberoamérica	2025	532 horas / 6.657.909 palabras

De los recursos lingüísticos a la capacidad tecnológica

La existencia de corpus amplios y diversos constituye una condición necesaria, pero no suficiente, para alcanzar una mayor autonomía tecnológica. Los datos solo se convierten en una infraestructura estratégica cuando existen instituciones capaces de conservarlos, documentarlos y hacerlos reutilizables; capacidad de cómputo para procesarlos; equipos especializados para desarrollar los modelos; y sistemas de evaluación que permitan conocer su rendimiento en las distintas variedades del español.

Entre 2022 y 2026, esta relación entre lengua, datos y tecnología adquirió una presencia creciente en las políticas públicas del espacio hispanohablante. El desarrollo de modelos en lenguas distintas del inglés comenzó a entenderse no solo como una cuestión de innovación, sino también como una forma de preservar capacidad científica, cultural y administrativa ante la concentración del mercado en unas pocas empresas internacionales.

En España, el PERTE Nueva Economía de la Lengua y la Estrategia Nacional de Inteligencia Artificial han articulado un conjunto de iniciativas que abordan diferentes componentes de esta cadena. Su relevancia no procede únicamente del número de proyectos impulsados, sino de la complementariedad entre la creación de recursos lingüísticos, el desarrollo de modelos fundacionales, la terminología especializada, la normalización, la observación y la producción de contenidos digitales. Como se puede observar, España ha configurado una política relativamente integral que combina financiación pública, desarrollo tecnológico, recursos lingüísticos, normalización idiomática, observación estratégica y proyección internacional de la lengua.

En el ámbito de los modelos fundacionales y los grandes modelos de lenguaje, iniciativas como ALIA, MarIA y BETO persiguen dotar al español de capacidades tecnológicas propias. Estos modelos permiten entrenar sistemas lingüísticos directamente sobre datos en español y en las lenguas cooficiales, por lo que reducen la necesidad de adaptar modelos inicialmente diseñados para el inglés. Además, contribuyen a mejorar la representación de fenómenos lingüísticos específicos del español y a poner en valor los recursos y conocimiento tecnológicos nacionales en un ámbito dominado por grandes empresas internacionales.

Estos modelos necesitan «alimentarse» de datos con los que puedan entrenarse, por ello la segunda línea estratégica es aumentar el número, calidad y disponibilidad de las infraestructuras lingüísticas en español y lenguas cooficiales mediante proyectos como ILENIA o CORPES XXI. Su relevancia es especialmente significativa porque la calidad de los modelos depende en gran medida de la cantidad, diversidad y representatividad de los datos utilizados durante el entrenamiento.

Tabla 6. Principales iniciativas en España para la contribución del español en IA
Fuente: elaboración propia a partir de datos del PERTE Nueva Economía de la Lengua (2022) y Estrategia Nacional de Inteligencia Artificial (2024)

Ámbito estratégico	Iniciativas principales	Contribución del español en IA
Contenidos digitales y recursos multimodales	**España Hub Audiovisual de Europa**	Generación de contenidos audiovisuales susceptibles de alimentar corpus multimodales y sistemas generativos. Impulso a la producción audiovisual en español para datos de IA.
Enseñanza y proyección internacional	**Instituto Cervantes Digital** (plataformas y recursos)	Internacionalización del español mediante plataformas digitales y tecnologías lingüísticas. Promoción del español en entornos IA, oralidad y contenidos educativos.
Infraestructuras lingüísticas y recursos de datos	**ILENIA** (Impulso de las Lenguas en IA) y **CORPES XXI** (RAE/ASALE)	Creación y puesta a disposición de corpus, datos y recursos lingüísticos reutilizables (texto, voz, multilingües). ILENIA coordina recursos para lenguas cooficiales; CORPES XXI: gran corpus de referencia del siglo XXI (cientos de millones de formas).
Investigación y supercomputación lingüística	**Barcelona Supercomputing Center (BSC)**	Infraestructura de entrenamiento de modelos lingüísticos (MareNostrum) y desarrollo de tecnologías avanzadas de PLN.
Modelos fundamentales y grandes modelos de lenguaje	**ALIA** (familia de modelos fundacionales abiertos), **MarIA** (BSC), **BETO** y derivados	Desarrollo de modelos de IA entrenados en español y lenguas cooficiales (catalán, gallego, euskera, etc.). ALIA incluye modelos como ALIA-40B, Salamandra (2B/7B/40B). MarIA: familia RoBERTa/GPT2 en español de BSC.
Normalización y calidad lingüística	**LEIA** (Lengua Española e Inteligencia Artificial, RAE)	Adaptación del español a asistentes virtuales, correctores, buscadores y sistemas conversacionales. Velar por el buen uso del español en máquinas e IA; herramientas lingüísticas basadas en IA para usuarios.
Observación, medición y prospectiva	**ONTSI, Observatorio Global del Español** (Instituto Cervantes)	Estudios sobre adopción de IA, diversidad lingüística y brecha digital con el inglés. Seguimiento del impacto de la IA sobre el español y análisis de tendencias.
Políticas públicas y financiación	**PERTE** Nueva Economía de la Lengua y **ENIA/ENIA+**	Financiación, gobernanza e impulso institucional del ecosistema lingüístico digital. Moviliza inversión pública y privada para toda la cadena de valor del español en IA, traducción, contenidos y investigación.
Terminología e IA para la ciencia y la traducción	**TeresIA**	Metabuscador panhispánico de terminología, extracción automática de términos y neologismos mediante IA, apoyo a traducción humana y automática, terminología científica y técnica en español. Genera corpus científico y valida términos.

Tabla 7. Principales iniciativas en Hispanoamérica para la contribución del español a la IA
Fuente: elaboración propia a partir de datos de Latam-GPT, Comisión Económica para América Latina (CEPAL), Centro Nacional de Inteligencia Artificial de Chile (CENIA), Organización para la Cooperación y el Desarrollo Económicos (OCDE) y Asociación de Academias de la Lengua Española (ASALE)

Ámbito estratégico	Iniciativas principales	Contribución del español a la IA	Países principales
Infraestructuras lingüísticas y recursos de datos	Contribuciones a corpus para Latam-GPT; iniciativas nacionales de datos abiertos	Creación y compartición de corpus multilingües para entrenar modelos más precisos y culturalmente relevantes.	Regional (coordinado por CENIA); México, Colombia y Chile
Investigación y supercomputación	CENIA (Chile); centros de IA nacionales; alianzas con AWS y otros	Infraestructura para entrenamiento de modelos en español y desarrollo de PLN regional.	Chile (CENIA) y México
Modelos de lenguaje	Latam-GPT (CENIA, Chile)	Primer modelo de lenguaje abierto y colaborativo entrenado con datos latinoamericanos (español y lenguas originarias).	Chile (líder), Colombia, México, Argentina, Perú, Ecuador, Uruguay y otros (17 países hispanohablantes)
Observación, medición y prospectiva	Índice Latinoamericano de Inteligencia Artificial (ILIA) (CENIA + CEPAL)	Monitoreo del impacto de la IA en el español y brechas lingüísticas/culturales en la región.	Regional (17 países hispanohablantes)
Políticas públicas y financiación	Estrategias nacionales de IA (varios países); Carta Iberoamericana de Inteligencia Artificial en la Administración Pública; apoyo CAF/BID	Marcos éticos, regulatorios y de gobernanza que priorizan datos en español y adopción inclusiva. Financiación multilateral para proyectos regionales.	Brasil, Chile, Colombia, México, Uruguay y Argentina
Terminología, normalización y calidad lingüística	Colaboraciones con ASALE; marcos éticos nacionales	Adaptación de terminología científica/técnica en español y herramientas de normalización para IA.	Varios (vía ASALE y estrategias nacionales)

Un papel singular corresponde a las iniciativas centradas en la terminología y la traducción especializada, entre las que destaca TeresIA. El proyecto responde a una de las necesidades más críticas de la inteligencia artificial contemporánea: la capacidad de comprender y generar conocimiento especializado en español. Mediante la extracción automática de terminología, la identificación de neologismos y la creación de recursos interoperables para la traducción científica y técnica, TeresIA contribuye a garantizar que el español mantenga su capacidad de expresión en ámbitos de alta especialización, particularmente en ciencia, tecnología e innovación.

La dimensión de la normalización lingüística está representada principalmente por el proyecto LEIA, promovido por la Real Academia Española. Su objetivo consiste en incorporar criterios lingüísticos normativos al desarrollo de asistentes virtuales, correctores automáticos, sistemas conversacionales y otras aplicaciones digitales. En un contexto en el que millones de usuarios interactúan diariamente con sistemas basados en inteligencia artificial, la forma lingüística se convierte en un elemento clave para preservar la cohesión y la unidad del espacio hispánico.

Por su parte, los organismos dedicados a la observación, medición y prospectiva, como el Observatorio Nacional de Tecnología y Sociedad (ONTSI) o el Observatorio Global del Español, desempeñan una función estratégica al proporcionar evidencia empírica sobre la evolución del ecosistema digital. Sus estudios permiten identificar tendencias, medir el grado de presencia del español en las tecnologías emergentes y orientar la toma de decisiones públicas y privadas.

La enseñanza y la proyección internacional del español, impulsadas por el Instituto Cervantes mediante sus procesos de digitalización, representan otra dimensión esencial. La incorporación de tecnologías lingüísticas avanzadas en la enseñanza del español amplía el alcance internacional de la lengua y genera nuevos recursos digitales susceptibles de integrarse en futuras aplicaciones de inteligencia artificial.

Finalmente, el desarrollo de contenidos digitales y recursos multimodales, promovido a través de programas como España Hub Audiovisual de Europa, adquiere una relevancia creciente en la era de los modelos multimodales. La producción de contenidos audiovisuales en español contribuye a aumentar la disponibilidad de datos de voz, imagen y vídeo, recursos cada vez más necesarios para entrenar sistemas capaces de procesar simultáneamente texto, audio e imágenes.

En conjunto, estas iniciativas configuran una estrategia que trasciende la mera digitalización de la lengua y se orienta hacia la construcción de una verdadera infraestructura lingüística para la inteligencia artificial. Su objetivo último no es únicamente incrementar la presencia cuantitativa del español en los sistemas de IA, sino garantizar que la diversidad

lingüística, cultural y terminológica de la comunidad hispanohablante se represente correctamente en las tecnologías que definirán la próxima fase de la transformación digital.

La estrategia de Hispanoamérica tiene un enfoque colaborativo hacia una IA más inclusiva y representativa del español de sus regiones, con Latam-GPT como emblema de soberanía regional frente al dominio de modelos anglocéntricos. Sus iniciativas destacan por integrar la diversidad cultural y lingüística —incluidas las variedades del español y las lenguas originarias—, aunque presentan brechas de financiación, infraestructura y ejecución con respecto a la estrategia española.

Las estrategias nacionales de IA (en países como Chile, Colombia y México), la Carta Iberoamericana de Inteligencia Artificial en la Administración Pública y el apoyo de organismos multilaterales, como el Banco de Desarrollo de América Latina y el Caribe (CAF) y el Banco Interamericano de Desarrollo (BID), fomentan una aproximación colaborativa y regional que prioriza la soberanía tecnológica y la inclusión lingüística. Esta combinación de inversión y coordinación latinoamericana fortalece aún más la presencia del español en los ecosistemas globales de IA. Gracias a ello, Hispanoamérica ha dado un salto con Latam-GPT, el primer gran modelo de lenguaje abierto desarrollado en la región (lanzado en 2026 por CENIA en Chile, con la participación de más de 15 países), entrenado con datos latinoamericanos que incorporan variedades regionales del español, culturas locales e incluso lenguas indígenas.

En colaboración con las academias que integran la ASALE, este esfuerzo se extiende por Hispanoamérica y contribuye a mantener la capacidad del español para nombrar las innovaciones científicas y tecnológicas. De este modo, se promueve una lengua actualizada, cohesionada y preparada para participar en el discurso científico internacional, al tiempo que se facilita el acceso al conocimiento en español a cientos de millones de hablantes.

Respecto a la observación, medición y prospectiva del impacto del español, el Índice Latinoamericano de Inteligencia Artificial (ILIA), elaborado por el Centro Nacional de Inteligencia Artificial (CENIA) y la Comisión Económica para América Latina (CEPAL), monitorea la adopción, gobernanza y brechas lingüísticas en 17 países hispanohablantes. Estos instrumentos prospectivos identifican oportunidades y riesgos, orientan políticas que posicionen al español como lengua central en la IA y anticipan su evolución en un mundo cada vez más automatizado. Como resultado de sus estudios y análisis, el ILIA revela un progreso en adopción, pero enfatiza sobre la necesidad de una mayor inversión y gobernanza. Esta aproximación regional fortalece la unidad panhispánica en la era digital, que posiciona al español como lengua competitiva en IA a nivel global. Por tanto, el ecosistema es prometedor pero aún incipiente, con un alto potencial de impacto en educación, gobierno y preservación lingüística y cultural.

La evolución de los modelos como resultado de esta infraestructura

La evolución reciente de los modelos en español permite analizar hasta qué punto los corpus y las políticas institucionales comienzan a traducirse en capacidades tecnológicas. En pocos años el ecosistema ha pasado de disponer principalmente de modelos de comprensión, orientados a clasificar textos, reconocer entidades o analizar sentimientos, a incorporar sistemas generativos, modelos especializados y desarrollos fundacionales abiertos de mayor escala.

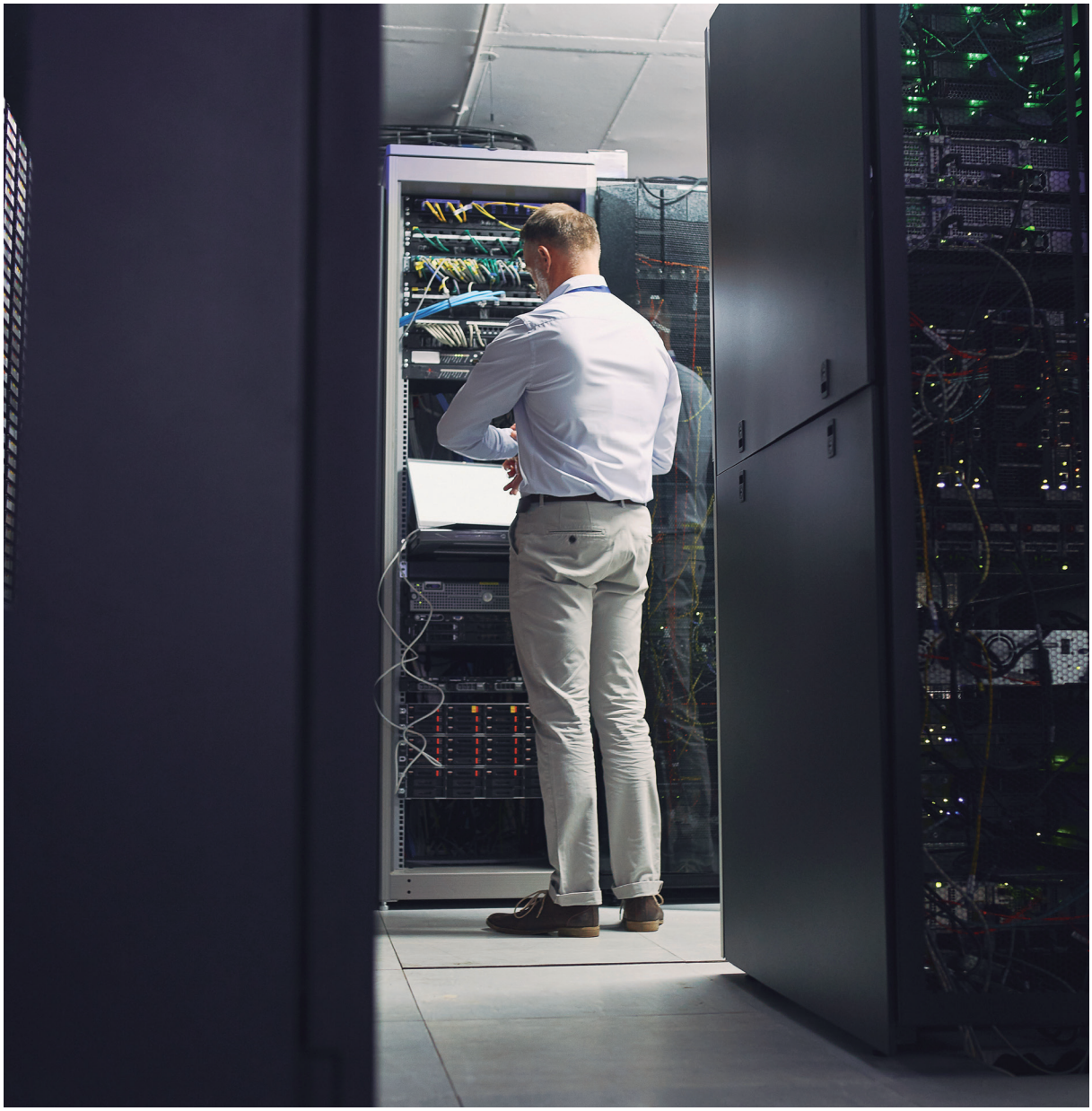


Tabla 8a. Modelos de lenguaje en español (2020-2026)
Fuente: elaboración propia a partir de las fichas técnicas y documentación oficial de los modelos en Hugging Face, Barcelona Supercomputing Center, PlanTL-GOB-ES, Universidad de Chile, IIC-UAM, Clibrain, Projecte AINA, ALIA y Latam-GPT

Modelo / familia	Institución / desarrollador	País / ámbito	Año de publicación	Idioma(s) principal(es)	Datos de entrenamiento / adaptación	Valor para la lengua española
BETO	Universidad de Chile / DCC	Chile / académico	2020	Español	Corpus amplio en español con Wikipedia, OPUS y otras fuentes; entrenamiento con Whole Word Masking.	Modelo pionero: demostró que un BERT monolingüe entrenado en español podía superar a modelos multilingües generales en tareas de comprensión.
MarIA / modelos BNE	Barcelona Supercomputing Center (BSC) / Plan de Tecnologías del Lenguaje	España / público-académico	2021	Español	570 GB de texto limpio y deduplicado; 135.000 millones de palabras del Archivo Web Español de la Biblioteca Nacional de España.	Supuso un salto de escala para el español al aprovechar el Archivo Web de la BNE y supercomputación pública.
BERTIN	BERTIN Project / comunidad Hugging Face	España / comunidad abierta	2022	Español	Entrenado sobre la parte española de mC4; aplica muestreo por perplejidad para reducir ruido y coste.	Aporta una aproximación data-centric al entrenamiento de modelos en español, útil para equipos con recursos computacionales limitados.
RigoBERTa	Instituto de Ingeniería del Conocimiento (IIC) / UAM-IIC	España / académico-empresarial	2022	Español	Entrenado sobre RigoCorpus, un corpus curado de subcorpus en español; adaptable a dominios como legal, financiero o salud.	Refuerza la línea de modelos discriminativos especializados en español y muestra la importancia de corpus curados frente a datos masivos sin depurar.
RoBERTuito	pysentimiento / comunidad académica	Argentina / académico	2021-2022	Español; también tareas con cambio de código inglés-español	Más de 500 millones de tuits en español.	Cubre registros informales, ortografía no estándar, emotividad y rasgos propios de la comunicación digital en español.
BERTuit	Universidad Politécnica de Madrid y colaboradores	España / académico	2022-2023	Español	Más de 230 millones de tuits en español.	Aporta un modelo nativo para el español informal de redes, con especial interés para estudios de desinformación.
Modelos biomédicos y clínicos en español	BSC / PlanTL-GOB-ES	España / biomédico-clínico	2021-2022	Español	Corpus biomédicos en español; CoWeSe; corpus clínico con más de 278.000 documentos y notas; 1.1B tokens biomédicos y 95M tokens EHR en trabajos asociados.	Muestran que el español también necesita modelos de dominio para preservar terminología y fraseología especializadas.

Tabla 8b. Modelos de lenguaje en español (2020-2026)
Fuente: elaboración propia a partir de las fichas técnicas y documentación oficial de los modelos en Hugging Face, Barcelona Supercomputing Center, PlanTL-GOB-ES, Universidad de Chile, IIC-UAM, Clibrain, Projecte AINA, ALIA y Latam-GPT

Modelo / familia	Institución / desarrollador	País / ámbito	Año de publicación	Idioma(s) principal(es)	Datos de entrenamiento / adaptación	Valor para la lengua española
MEL	Instituto de Ingeniería del Conocimiento (IIC)	España / jurídico	2025	Español jurídico	Corpus de documentos jurídicos españoles, incluidos BOE, textos parlamentarios, sentencias y otros textos legislativos.	Actualiza el ecosistema legal en español con un modelo especializado posterior a RoBERTalex y orientado a tareas de comprensión jurídica.
MrBERT-es	BSC-LT	España / europeo	2026	Español e inglés	Adaptación de vocabulario desde MrBERT; familia MrBERT preentrenada sobre 35 lenguas europeas y código.	Introduce arquitecturas encoder modernas con contexto largo para español, útiles en producción y recuperación semántica.
MrBERT de dominio	BSC-LT	España / europeo	2026	Multilingüe con peso relevante del español e inglés	MrBERT-legal: corpus legal de 8B tokens; MrBERT-biomed: 24,13B tokens biomédicos; MrBERT-science: 3,6B tokens.	Señalan la evolución desde modelos generales hacia modelos de dominio capaces de procesar ciencia, derecho y biomedicina en español.
LINCE	Clibrain	España / empresarial	2023	Español	LINCE-ZERO se basa en Falcon-7B y se ajusta con un conjunto propietario de 80.000 ejemplos de instrucciones en español; variantes posteriores adaptan Llama 2 y Mistral.	Representa la entrada de actores privados españoles en modelos generativos orientados específicamente a instrucciones en español.
FLOR	Projecte AINA / BSC	España / catalán-español-inglés	2023-2024	Catalán, español e inglés	FLOR-6.3B se obtiene por adaptación de vocabulario y preentrenamiento continuo con 140B tokens.	Aunque su foco principal es el catalán, incorpora español como lengua de trabajo y demuestra la utilidad de modelos plurilingües ibéricos.
Ăguila-7B	Projecte AINA / BSC	España / catalán-español-inglés	2022-2023	Catalán, español e inglés	Entrenado con un corpus trilingüe de 26B tokens procedente de corpus públicos y rastreadores.	Refuerza la línea de modelos abiertos para lenguas de España y muestra la convergencia entre español y lenguas cooficiales en la IA.
Salamandra / ALIA	BSC / Gobierno de España / ALIA	España / europeo	2024-2025	35 lenguas europeas, español, catalán, gallego, euskera y código	Datos multilingües altamente curados; ALIA-40b indica 9,37 billones de tokens; Salamandra se entrena desde cero con datos abiertos y documentación técnica.	Amplía el valor del español reciente del ecosistema español: pasa de modelos de comprensión a modelos generativos fundacionales abiertos de escala decenas de miles de millones.

Modelo / familia	Institución / desarrollador	País / ámbito	Año de publicación	Idioma(s) principal(es)	Datos de entrenamiento / adaptación	Valor para la lengua española
Salamandra TA	BSC-LT	España / traducción automática	2025	Lenguas europeas; foco en catalán, español e inglés en algunas variantes	Preentrenamiento continuo de Salamandra con datos monolingües y paralelos; ajuste supervisado con instrucciones de traducción.	Amplía el valor del español dentro de sistemas multilingües de traducción abierta, especialemente en el espacio europeo e ibérico.
Latam-GPT	CENIA, Gobierno de Chile, CAF, AWS, Data Observatory y red regional	América Latina y Caribe	2026	Español latinoamericano, portugués y lenguas indígenas en el proyecto	Datos latinoamericanos seleccionados y documentados; enfoque regional, abierto y colaborativo.	Es la iniciativa más relevante para corregir la subrepresentación de variedades latinoamericanas y reforzar la soberanía tecnológica regional.

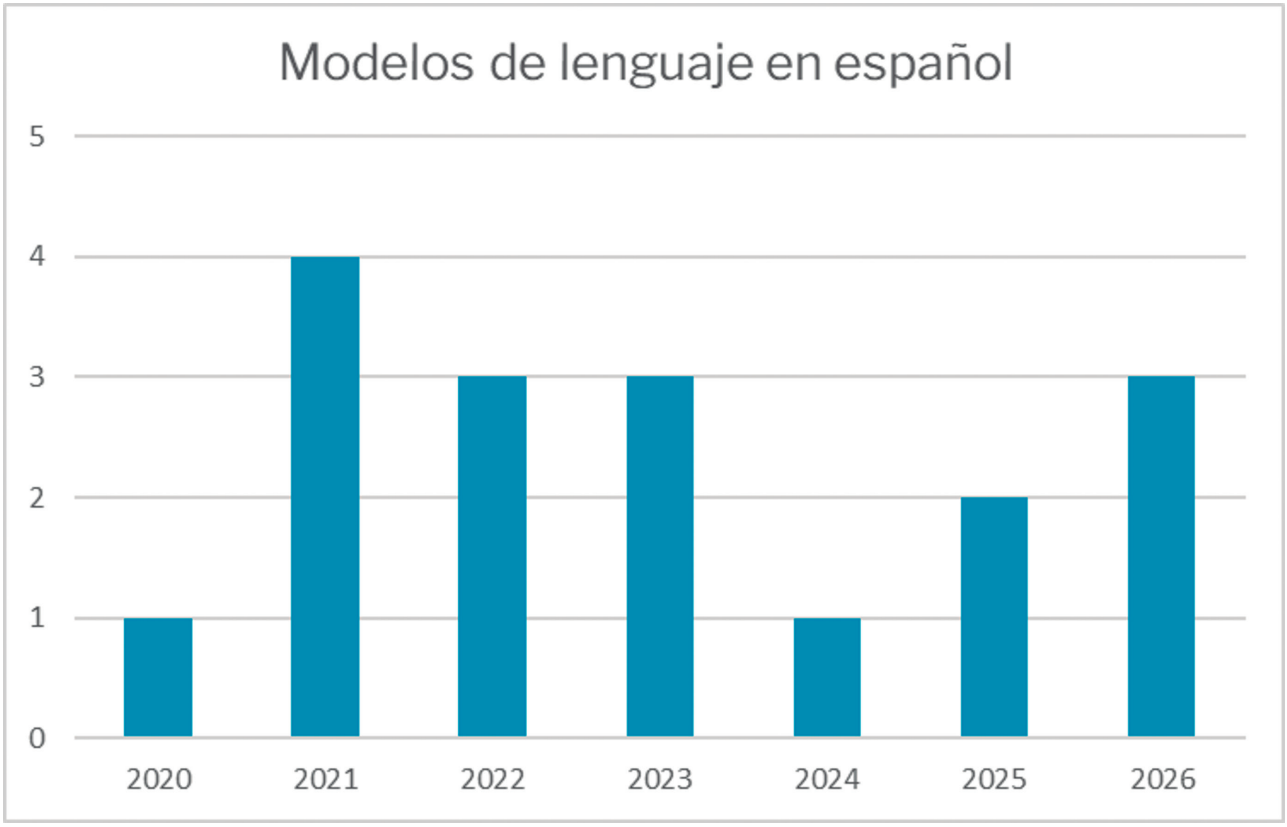
Los primeros modelos se centraron fundamentalmente en la comprensión automática del español. BETO, desarrollado por la Universidad de Chile, fue uno de los primeros modelos basados en BERT entrenados específicamente para esta lengua y se convirtió en una referencia para distintas tareas de procesamiento del lenguaje natural. MarIA, impulsado por el Barcelona Supercomputing Center, supuso un salto de escala al utilizar el Archivo de la Web Española de la Biblioteca Nacional de España para entrenar modelos abiertos en español. BERTIN y RigoBERTa, por su parte, reforzaron la importancia de la calidad, la selección y la depuración de los datos de entrenamiento. A esta línea se sumaron modelos orientados a registros específicos, como RoBERTuito y BERTuit, entrenados con grandes volúmenes de textos procedentes de redes sociales. Su relevancia radica en su capacidad para representar usos informales, digitales y conversacionales del español, escasamente presentes en los corpus normativos o institucionales. Estos modelos resultan especialmente útiles para analizar la opinión pública, el discurso digital, la desinformación, la polarización y el lenguaje espontáneo en los entornos sociales.

Otra tendencia destacada es la especialización por ámbitos de conocimiento. En los últimos años han aparecido modelos adaptados al lenguaje biomédico, clínico, jurídico y científico, entre ellos los desarrollados por el BSC y PlanTL-GOB-ES, así como recursos como RoBERTalex, MEL y MrBERT-es y sus variantes especializadas. Estos modelos muestran que la presencia del español en la IA no depende únicamente de los modelos generales, sino también de herramientas capaces de procesar con precisión la terminología, la documentación técnica y los textos altamente especializados. Su desarrollo resulta estratégico para sectores como la salud, la justicia, la Administración pública, la investigación y la economía del conocimiento. La etapa más reciente está marcada por la aparición de modelos

generativos y fundacionales. Iniciativas como LINCE, FLOR, Ăguila, Salamandra y ALIA reflejan la transición hacia modelos abiertos, plurilingües y de mayor escala. En particular, Salamandra y ALIA representan una apuesta institucional por dotar al español y a las lenguas cooficiales de infraestructuras propias de IA generativa. Estos modelos contribuyen a reducir la dependencia de sistemas desarrollados principalmente en inglés y favorecen la creación de aplicaciones adaptadas a las necesidades lingüísticas, culturales y administrativas del espacio hispanohablante.

En paralelo, Latam-GPT incorpora una dimensión especialmente relevante para la internacionalidad del español. Coordinado desde Hispanoamérica, este proyecto busca integrar datos regionales y representar con mayor precisión las variedades lingüísticas y las realidades culturales y sociales latinoamericanas. Su importancia radica en que responde a uno de los principales desafíos actuales: conseguir que los modelos de lenguaje reflejen la diversidad del mundo hispanohablante y representen de manera equilibrada sus distintas variedades.

Gráfico 5. Número de modelos en español que se han desarrollado anualmente desde 2020 a 2026
Fuente: elaboración propia a partir de las fichas técnicas y documentación oficial de los modelos en Hugging Face, Barcelona Supercomputing Center, PlanTL-GOB-ES, Universidad de Chile, IIC-UAM, Clibrain, Projecte AINA, ALIA y Latam-GPT



En conjunto, el panorama de los modelos de lenguaje en español muestra una evolución en cuatro direcciones complementarias. La primera es la consolidación de modelos monolingües de comprensión, como BETO, MarIA, BERTIN y RigoBERTa. La segunda corresponde a la especialización por registros y ámbitos de conocimiento, con modelos destinados a las redes sociales, la salud, el derecho y la ciencia. La tercera es la aparición de modelos generativos abiertos y plurilingües, como LINCE, FLOR, Ăguila, Salamandra y ALIA. La cuarta es el desarrollo de una respuesta latinoamericana articulada en torno a Latam-GPT.

No obstante, esta evolución debe interpretarse con cautela. No todos los modelos existentes tienen la misma naturaleza ni el mismo grado de soberanía tecnológica. Algunos han sido entrenados desde cero con corpus propios; otros son adaptaciones, ajustes finos o especializaciones de modelos internacionales. Algunos están orientados a comprender textos; otros, a generar lenguaje; y otros, a recuperar información, clasificar documentos o apoyar sistemas de búsqueda semántica. Esta diversidad representa una madurez en el entorno de la IA: el español empieza a disponer de un ecosistema de modelos más amplio, capaz de cubrir registros, dominios, variedades y necesidades tecnológicas distintas.

El reto de los próximos años será convertir esta variedad de iniciativas en una infraestructura estable, interoperable y evaluable. Para ello será necesario reforzar los corpus de entrenamiento, documentar y etiquetar mejor los datos, ampliar la representación de variedades americanas, consolidar bancos de evaluación nativos en español y favorecer la reutilización abierta de modelos por parte de universidades, administraciones, empresas y comunidades de desarrolladores. Solo así el español podrá avanzar desde la disponibilidad de modelos aislados hacia una verdadera soberanía lingüística en inteligencia artificial.



5. Nuevos desafíos y líneas de actuación para el español en la inteligencia artificial



La construcción de corpus, infraestructurasy modelos propios no responde de manera definitiva a la cuestión de la presencia del español en la inteligencia artificial. Al contrario, abre una nueva etapa en la que el desafío ya no consiste únicamente en aumentar el volumen de los recursos disponibles, sino en garantizar su calidad, diversidad y aprovechamiento efectivo. La posición futura de la lengua dependerá de la capacidad para coordinar datos, conocimiento lingüístico, infraestructura tecnológica, evaluación independiente y políticas públicas sostenidas.

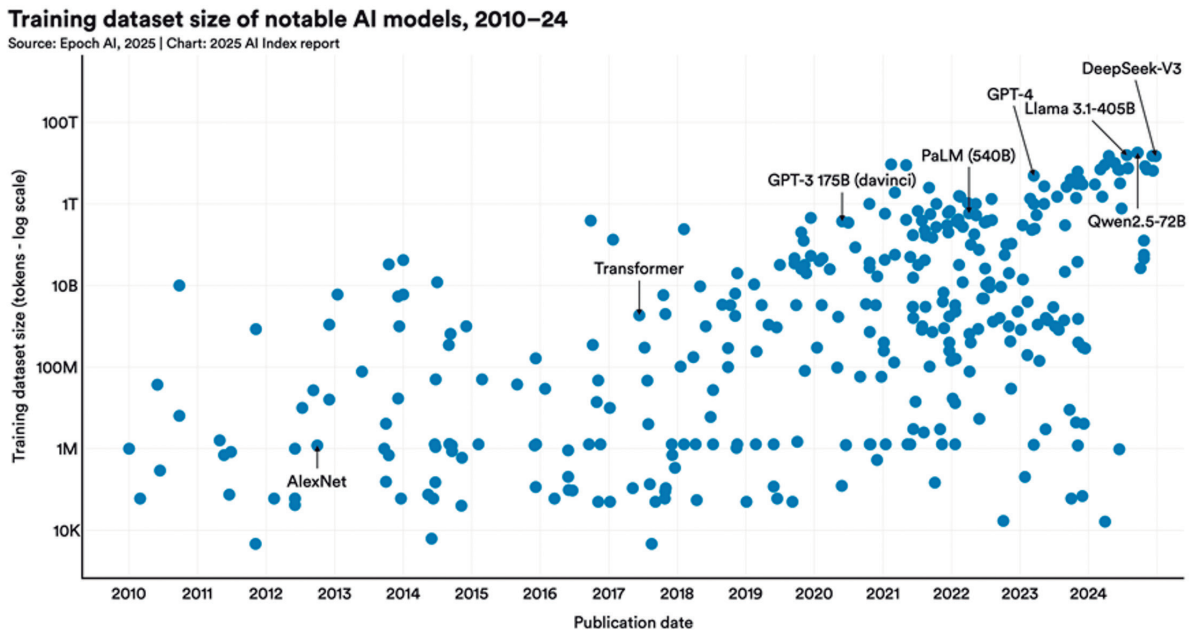
Escala, calidad y trazabilidad de los datos

La irrupción de los grandes modelos fundacionales ha modificado las condiciones tradicionales de compilación de corpus. Mientras los recursos lingüísticos clásicos se diseñaron para garantizar representatividad, equilibrio documental, trazabilidad de las fuentes y control experto, los modelos actuales requieren conjuntos de datos de una escala sin precedentes, obtenidos con frecuencia mediante procesos automatizados de rastreo web. El *AI Index Report* 2025 de la Universidad de Stanford señala que el tamaño de los conjuntos utilizados para entrenar grandes modelos de lenguaje se duplica aproximadamente cada ocho meses, lo que refleja la aceleración de la demanda de datos en la nueva etapa de la IA generativa.

Este aumento de escala obliga a reconsiderar cuestiones que antes pertenecían principalmente al ámbito de la lingüística de corpus, pero que hoy se sitúan en el centro de la gobernanza de la inteligencia artificial: la calidad de los datos, los derechos de autor, la protección de la información personal, la trazabilidad de las fuentes, la representación de los distintos grupos sociales y la documentación de los procesos de recopilación y depuración.

En julio de 2025, la Comisión Europea publicó una plantilla común para que los proveedores de modelos de propósito general presenten resúmenes públicos de los contenidos utilizados en su entrenamiento. Esta obligación, vinculada al artículo 53 del Reglamento de Inteligencia Artificial, establece una base mínima de transparencia sobre las fuentes y los procedimientos de tratamiento de los datos. No permite reconstruir íntegramente los corpus, pero facilita el ejercicio de derechos y favorece un conocimiento más preciso de los materiales con los que se desarrollan los modelos.

Gráfico 6. Tamaño en tokens de los conjuntos de datos de entrenamiento de los grandes modelos de IA
Fuente: Stanford HAI, AI Index Report 2025



Para el español, avanzar en esta dirección significa dotar a los corpus y conjuntos de datos de una documentación lingüística más detallada. No basta con asignarles una etiqueta general como «español» o «Spanish»: es necesario describir su procedencia territorial, cronología, modalidad, registro, ámbito temático, perfil de los hablantes y condiciones de uso (Williams, Nangia y Bowman, 2018).

A este problema se añade la utilización creciente de contenidos generados por inteligencia artificial. Si estos materiales vuelven a incorporarse sin identificar a los corpus de entrenamiento, pueden producirse procesos circulares que amplifiquen errores, reduzcan la diversidad expresiva o favorezcan la reproducción de formas ya dominantes. La trazabilidad deberá incluir, por tanto, no solo el origen geográfico y documental de los datos, sino también la distinción entre contenidos humanos, sintéticos y elaborados mediante procedimientos mixtos.

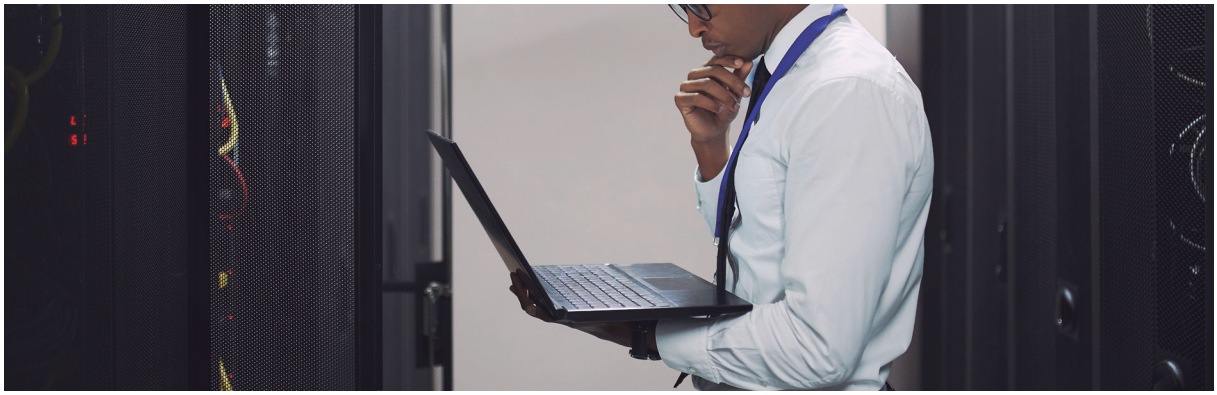
La representación del español pluricéntrico

El segundo gran desafío es la representatividad. El español dispone de una extensión internacional considerable, pero esa diversidad geográfica no se traslada automáticamente a los datos digitales. Las comunidades con mayor conectividad, producción editorial o presencia institucional generan más contenidos rastreables, mientras que otras variedades, grupos sociales y tradiciones discursivas quedan menos documentados.

El riesgo no es únicamente que los modelos cometan errores ante determinados acentos o usos regionales. Una representación desequilibrada puede llevarlos a interpretar como generales formas propias de unos pocos territorios, a corregir expresiones legítimas de otras variedades o a asociar ciertos usos lingüísticos con valoraciones sociales negativas. La calidad de un modelo en español no debería medirse solo por su corrección gramatical, sino también por su capacidad para reconocer la variación sin convertirla automáticamente en desviación.

En este sentido, PRESEEA ofrece un referente especialmente valioso. Su recopilación homogénea de entrevistas en numerosas comunidades urbanas permite disponer de muestras comparables de habla real procedentes de distintos países, generaciones, grupos sociales y niveles educativos. Esta dimensión panhispánica puede contribuir al entrenamiento y, sobre todo, a la evaluación de sistemas de reconocimiento del habla y modelos conversacionales más sensibles a la diversidad sociolingüística del idioma.

La mejora de la representación exige ampliar la participación de instituciones latinoamericanas en la selección, anotación y gobernanza de los datos. No se trata únicamente de incorporar más documentos procedentes de América, sino de integrar conocimiento local sobre sus variedades, contextos culturales y necesidades sociales. La construcción de una infraestructura hispánica requiere criterios compartidos, pero también una gestión distribuida que evite reproducir nuevas formas de centralización lingüística.



Oralidad, multimodalidad y ámbitos especializados

Un tercer desafío procede del avance hacia modelos capaces de combinar texto, voz, imagen y vídeo. El análisis de Hugging Face muestra que el español dispone de una base apreciable en traducción y tecnologías del habla, pero mantiene una presencia muy reducida en tareas como texto a imagen y texto a vídeo. Esta debilidad resulta especialmente relevante porque buena parte de la interacción futura con la inteligencia artificial será oral y multimodal.

Los corpus escritos han alcanzado un grado considerable de madurez, mientras que los recursos orales continúan siendo más escasos, heterogéneos y costosos de producir. La recopilación de voz exige consentimiento, anonimización, transcripción, alineación temporal y anotación especializada. También requiere representar edades, acentos, condiciones acústicas y situaciones comunicativas diversas. Sin este esfuerzo, las tecnologías conversacionales pueden funcionar adecuadamente para unos hablantes y ofrecer resultados mucho menos precisos para otros.

Recursos como el Corpus Español Multimodal de Actos de Habla (COR.E.M.A.H.) anticipan las necesidades de esta nueva etapa. La integración de vídeos, transcripciones, gestos, información prosódica y etiquetado pragmático permite estudiar aspectos de la comunicación que no pueden deducirse únicamente de las palabras. Su interés reside en mostrar que comprender una lengua implica también interpretar la intención, la entonación, el contexto y la relación entre la expresión verbal y el comportamiento no verbal.

Junto con la multimodalidad, será necesario reforzar los datos pertenecientes a ámbitos especializados. La presencia del español en la inteligencia artificial dependerá también de su capacidad para expresar y procesar conocimientos científicos, médicos, jurídicos, técnicos y administrativos. Los corpus terminológicos y sectoriales, acompañados de metadatos rigurosos y sistemas de actualización, resultan indispensables para evitar que el idioma quede limitado a los usos generales mientras el conocimiento especializado se procesa principalmente en inglés.



Evaluar los modelos desde el español

La disponibilidad de corpus y modelos debe ir acompañada de sistemas de evaluación concebidos desde la propia lengua. Una puntuación global puede ocultar diferencias importantes entre países, registros, tareas y grupos de hablantes. Un sistema que funciona correctamente en textos periodísticos peninsulares puede mostrar resultados mucho menos satisfactorios ante una conversación caribeña, un documento administrativo hispanoamericano o una consulta médica formulada en lenguaje coloquial.

Por ello, resulta necesario construir bancos de evaluación hispánicos que midan de forma diferenciada la comprensión, la generación, la traducción, el reconocimiento de voz y las capacidades multimodales. Estos bancos deberían incluir variedades territoriales y sociales, registros formales e informales, terminología especializada y fenómenos pragmáticos. También deberían evaluar errores que afectan específicamente a la lengua: neutralización indebida de variedades, recomendaciones normativas incorrectas, traducciones literales, invención de términos o reproducción de estereotipos culturales.

La evaluación tampoco puede limitarse a comprobar si una respuesta resulta fluida; los grandes modelos pueden generar textos gramaticalmente correctos y, al mismo tiempo, ofrecer información falsa, simplificar diferencias culturales o atribuir a una variedad usos que no le corresponden. La participación de lingüistas, terminólogos, traductores y especialistas en ámbitos de conocimiento específicos resulta, por tanto, imprescindible para complementar las métricas automáticas con procedimientos de evaluación experta.

Los efectos de la inteligencia artificial sobre la propia lengua

La relación entre el español y la inteligencia artificial no se reduce a la forma en que los modelos aprenden el idioma. Estos sistemas producen ya cantidades crecientes de textos, traducciones, resúmenes, imágenes y contenidos audiovisuales que vuelven a circular en el espacio público. La IA no solo representa la lengua: también interviene en su producción y difusión.

Esta capacidad plantea un desafío nuevo. Si unas pocas herramientas generan una parte creciente de los contenidos digitales, sus preferencias léxicas, sintácticas y discursivas pueden extenderse entre los

usuarios y favorecer cierta homogeneización expresiva. Las variedades menos representadas podrían perder visibilidad frente a un español aparentemente neutro, pero construido a partir de distribuciones de datos y decisiones empresariales poco transparentes.

Será necesario observar cómo influyen estos sistemas en la escritura profesional, la educación, los medios de comunicación, la traducción y la administración pública. Esta vigilancia no debe orientarse a impedir la innovación, sino a detectar procesos de empobrecimiento, uniformización o dependencia antes de que se consoliden. El seguimiento de los usos lingüísticos generados o mediados por IA deberá convertirse en una línea estable de investigación.

De los proyectos a una infraestructura hispánica estable

Por último, el principal desafío institucional consiste en transformar la suma de corpus, modelos e iniciativas en una infraestructura estable. Muchos recursos lingüísticos dependen de proyectos temporales o nacionales, y pierden capacidad operativa cuando termina su financiación. La creación inicial debe ir acompañada de mecanismos de mantenimiento, actualización, documentación y preservación a largo plazo.

Esta infraestructura requiere inversión pública sostenida, capacidad de cómputo, equipos interdisciplinares y condiciones claras de reutilización. También exige coordinar las actuaciones de todos los países hispanohablantes, de modo que los recursos puedan compartirse sin perder la trazabilidad ni el control sobre sus usos. Esta apertura debe entenderse como una condición que favorece la innovación, pero ha de conciliarse con los derechos de autor, la protección de datos y el reconocimiento de las comunidades que generan los materiales.

Desde esta perspectiva, las principales líneas de mejora pueden resumirse en seis actuaciones: ampliar los corpus orales y multimodales; reforzar la representación de las variedades americanas y de los grupos menos documentados; adoptar estándares comunes de metadatos y documentación; crear bancos hispánicos de evaluación; garantizar financiación e infraestructura de cómputo estables; y establecer mecanismos independientes de seguimiento sobre los efectos lingüísticos y sociales de los modelos.

El español no parte de una situación de carencia total. Cuenta con una comunidad de usuarios de gran tamaño, una tradición consolidada de investigación lingüística, corpus de referencia y un ecosistema creciente de modelos e iniciativas públicas. El reto consiste ahora

en articular esos activos dentro de una estrategia compartida. La posición futura de la lengua en la inteligencia artificial dependerá menos de acumular datos de manera indiscriminada que de la capacidad para convertirlos en recursos de calidad, representativos y gobernados desde el conjunto del espacio hispanohablante.



6. Conclusiones

El análisis conjunto de la adopción, los datos, los corpus y los modelos permite concluir que el español está dejando de ser únicamente una lengua de uso de la inteligencia artificial para convertirse también en una lengua de desarrollo tecnológico. La demanda social, la disponibilidad de grandes colecciones textuales y la aparición de modelos propios configuran una base favorable para ampliar su capacidad computacional.

La posición del idioma presenta, sin embargo, distintos grados de madurez. El español dispone de una presencia consolidada en contenidos web, corpus textuales, traducción y tecnologías de la voz. Al mismo tiempo, comienza a incorporarse a los modelos generativos y cuenta con un amplio margen de crecimiento en imagen, vídeo y otros entornos multimodales. Esta distribución permite identificar con mayor precisión sus fortalezas y orientar los próximos esfuerzos hacia las áreas menos desarrolladas.

Uno de los principales activos del español reside en la combinación de escala y diversidad. Su carácter pluricéntrico puede convertirse en una ventaja para crear sistemas capaces de operar en contextos lingüísticos y culturales muy diferentes. Para aprovechar este potencial, la diversidad panhispánica debe incorporarse como un criterio técnico desde la selección y documentación de los datos hasta la evaluación final de los modelos.

El desafío principal ya no consiste tanto en la escasez absoluta de recursos como en su articulación. Existen corpus de referencia, conocimiento lingüístico especializado, instituciones consolidadas, iniciativas públicas y una comunidad creciente de desarrolladores. La prioridad es conectar estos elementos mediante estándares comunes, licencias claras, capacidad de cómputo, financiación estable y mecanismos compartidos de evaluación.

Desde esta perspectiva, la soberanía tecnológica del español será necesariamente cooperativa. La complementariedad entre España y los países hispanoamericanos permite reunir capacidades institucionales, recursos lingüísticos y variedades que difícilmente podrían representarse desde estrategias exclusivamente nacionales. Iniciativas como ALIA, Salamandra o Latam-GPT apuntan hacia una infraestructura distribuida en la que la cooperación panhispánica puede convertirse en una ventaja tecnológica.

A ello se suma una cuestión que adquirirá una importancia creciente: la inteligencia artificial no solo aprende el español, sino que comienza a intervenir en su evolución. Los textos, traducciones y contenidos generados por estos sistemas vuelven a circular entre los hablantes y pueden influir en sus preferencias lingüísticas. Observar estos efectos será

tan importante como medir la presencia del idioma en los datos y los modelos.

El español se encuentra, en definitiva, ante una oportunidad favorable. Cuenta con hablantes, demanda, datos, instituciones y capacidad científica. El siguiente paso consiste en transformar estos activos en una infraestructura panhispánica estable, abierta y evaluable que permita consolidar la lengua como espacio de comunicación, conocimiento, innovación y creación de valor.

Bibliografía

ALIA (s. f.). La *infraestructura pública de inteligencia artificial en castellano y lenguas cooficiales*. https://alia.gob.es/

Bender, E. M. y Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics, 6,* 587–604. https://doi.org/10.1162/tacl_a_00041

Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H. y Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. En *Proceedings of the Practical ML for Developing Countries Workshop at ICLR 2020*. https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf

Comisión Europea (24 de julio de 2025). *Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models*. https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models

Common Crawl (s. f.). *Statistics of Common Crawl Monthly Archives. Primary language distribution*. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

Consejo Superior de Investigaciones Científicas (s. f.). *TeresIA: Terminología más inteligencia artificial*. Proyecto TeresIA. https://proyectoteresia.org/

Gobierno de España (2022). *PERTE Nueva economía de la lengua. Memoria técnica*. https://planderecuperacion.gob.es/sites/default/files/2022-03/PERTE_Nueva_economia_de_la_lengua_memoria_16032022.pdf

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

High Performance Language Technologies (s. f.). *HPLT Datasets v2.0*. https://hplt-project.org/datasets/v2.0

Hugging Face (s. f.). *Models*. https://huggingface.co/models

Instituto Cervantes y Ministerio de Asuntos Económicos y Transformación Digital (2023). *Informe del estado actual de los corpus en español, lenguas cooficiales y variantes del español*. https://cervantes.org/es/sobre-nosotros/publicaciones/informe-estado-actual-corpus-espanol-lenguas-cooficiales-variantes

LatamGPT (s. f.). *LatamGPT: Inteligencia artificial para América Latina y el Caribe*. https://www.latamgpt.org/

Microsoft AI Economy Institute (2026). *Global AI diffusion: Q1 2026 trends and insights*. Microsoft. https://www.microsoft.com/en-us/research/wp-content/uploads/2026/05/Microsoft-AI-Diffusion-Report-2026-Q1.pdf

PlanTL-GOB-ES (s. f.). *Spanish language models and resources* [Repositorio de GitHub]. GitHub. https://github.com/PlanTL-GOB-ES/lm-spanish

Powell, P. y Smalley, I. (2025). *¿Qué es la deduplicación de datos?* IBM. https://www.ibm.com/es-es/think/topics/data-deduplication

Proyecto para el Estudio Sociolingüístico del Español de España y de América (PRESEEA) (2014). *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América* [Corpus]. Universidad de Alcalá. https://preseea.uah.es/corpus-preseea

Real Academia Española (s. f.). *Lengua Española e Inteligencia Artificial (LEIA)*. https://www.rae.es/leia-lengua-espanola-e-inteligencia-artificial

Real Academia Española (2026). *Corpus del Español del Siglo XXI (CORPES XXI), versión 1.5* [Corpus]. https://www.rae.es/banco-de-datos/corpes-xxi

Real Academia Española (29 de junio de 2026). *El CORPES supera los 455 millones de formas*. https://www.rae.es/noticia/el-corpes-supera-los-455-millones-de-formas

Soto, Á., Durán, R., Moreno, A., Adasme, S., Rovira, S., Jordán, V. y Poveda, L. (Coords.) (2025). *Índice Latinoamericano de Inteligencia Artificial (ILIA) 2025. Documentos de Proyectos* (LC/TS.2025/68/Rev.1). Comisión Económica para América Latina y el Caribe y Centro Nacional de Inteligencia Artificial. https://www.cepal.org/es/publicaciones/82514-indice-latinoamericano-inteligencia-artificial-ilia-2025

Stanford Institute for Human-Centered Artificial Intelligence (2025). *The 2025 AI Index Report*. Stanford University. https://hai.stanford.edu/ai-index/2025-ai-index-report

TodoELE (s. f.). *COR.E.M.A.H.: Corpus Español Multimodal de Actos de Habla*. https://todoele.net/corpus/coremah-corpus-espanol-multimodal-de-actos-de-habla

Williams, A., Nangia, N. y Bowman, S. R. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. En M. Walker, H. Ji y A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers),* 1112–1122. Association for Computational Linguistics.

Observatorio Global del Español
Instituto Cervantes

Informes del Observatorio Global del Español

NIPO: 110-26-002-0
ISSN: 3101-4267