Lecciones urgentes de español para la inteligencia artificial

La escasa inversión en su adaptación y la ventaja del inglés como lengua innovadora por excelencia relegan al castellano a un papel secundario en esta tecnología

LAURA MONTERO CARRETERO

El español es la segunda lengua materna del mundo tras el chino mandarín y cuenta con más de 585 millones de hablantes en todo el mundo, una realidad lingüística que contrasta con su escaso protagonismo como base para entrenar a los algoritmos de la inteligencia artificial (IA). Bajo esta etiqueta se engloban diferentes tecnologías, cada una de las cuales precisa un tipo de datos para funcionar. Las de imagen se nutren de los píxeles, las del lenguaje tienen como punto de partida el texto escrito y las de voz procesan las ondas sonoras para después convertirlas a texto, pero el denominador común a todas ellas es que necesitan aprender y, hasta ahora, lo han hecho de forma mayoritaria en inglés, siendo este el idioma en el que mejor funcionan. El castellano, por su parte, continúa como el eterno aspirante, ya que aún representa menos del 30% del mercado mundial de las tecnologías de procesamiento de lenguaje natural.

El uso de este activo como lengua nativa de la IA es una mina de oro apenas explotada con la que España conseguiría una ventaja competitiva importante y recortaría distancias a Estados Unidos y China, actuales líderes en la batalla por el dominio de esta tecnología llamada a cambiar para siempre la economía global.

El retraso frente al inglés no es atribuible a una única causa. Obedece, por un lado, a una razón histórica, pues las teorías pioneras en torno a esta disciplina surgieron en los años cincuenta del siglo XX en Estados Unidos, vinculadas a figuras como Marvin Minsky o Alan Turing. En nuestros días, se ha consolidado como la lengua de la investigación por excelencia. «Llama la atención que haya un mayor número de hablantes nativos del español que del inglés y sea este último donde más foco se pone. Es porque la comunidad científica toma como referencia el inglés», afirma Ricardo Moya, experto en inteligencia artificial de IMF.

La lengua de Shakespeare copa, por consiguiente, los principales desarrollos y publicaciones científicas. «La mayor parte de los entrenamientos y de los sistemas más potentes se han realizado desde universidades americanas, aunque esta tendencia está cambiando en los últimos años porque el sudeste asiático, y en particular China, está tomando una relevancia impresionante tanto en investigación como en empresas», explica Elena González-Blanco, directora general de Coverwallet en Europa y experta en inteligencia artificial y tecnología lingüística.

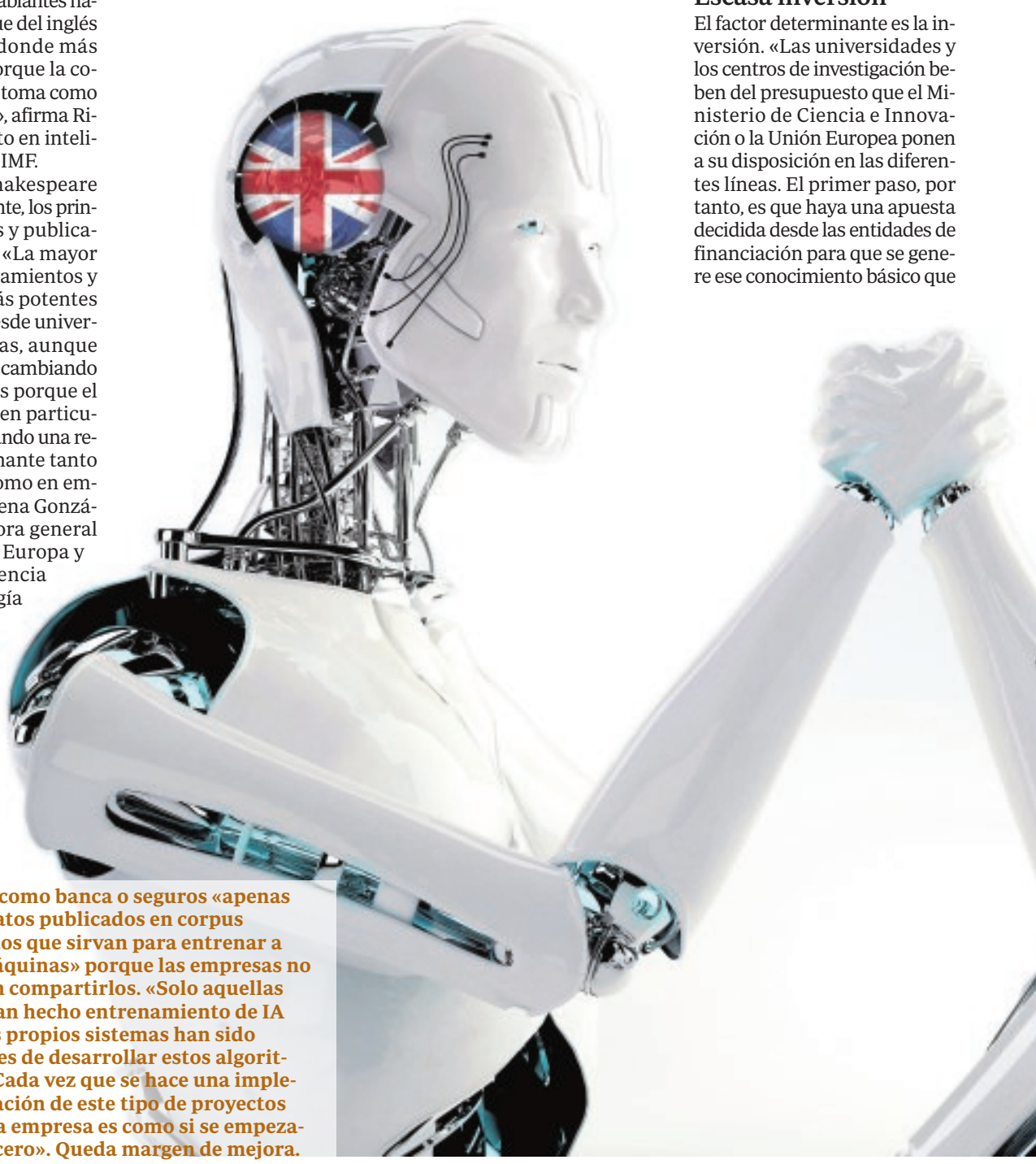
Por otro lado, las grandes tecnológicas –Google, Facebook, Amazon y Microsoft–, son estadounidenses. «Tienen una cantidad de datos más grande que nadie y cuentan con la capacidad para explotarlos. El resultado es que en áreas específicas de la IA, como la parte del lenguaje, hay unas diferencias enormes del funcionamiento de estas máquinas en inglés y en el resto de las lenguas», añade.

En el caso del español se suma la peculiaridad de sus variedades en América Latina, un aspecto en el que, tal y como señala la experta, queda camino por recorrer. «No se han hecho apenas trabajos para entrenar a las máquinas con las distintas variedades de español. De hecho, empresas de nuestro país que expanden allí su negocio con soluciones de lenguaje se encuentran con estas dificultades», cuenta. Sí se han concentrado esfuerzos, en cambio, en las lenguas peninsulares. «En España –dice– se ha hecho bastante en temas de traducción automática al gallego, euskera o catalán y existen empresas especializadas en ello».

Las fuentes consultadas coinciden en que, en cualquier caso, las características propias del español no son un escollo para su introducción en el ámbito de la inteligencia artificial. «Otros idiomas, como el inglés, usan menos 'stop words' (preposiciones, artículos...), pero todos tienen un vocabulario muy rico y aceptan expresiones complejas. Si desde el principio se hubieran destinado los mismos recursos al español que al inglés, la situación actual sería perfectamente inversa», considera Luis de la Fuente, investigador de la Universidad Internacional de la Rioja (Unir).

Escasa inversión

El factor determinante es la inversión. «Las universidades y los centros de investigación beben del presupuesto que el Ministerio de Ciencia e Innovación o la Unión Europea ponen a su disposición en las diferentes líneas. El primer paso, por tanto, es que haya una apuesta decidida desde las entidades de financiación para que se genere ese conocimiento básico que



PARTIR DE CERO

El 18% de las grandes empresas españolas –más de 250 empleados– han incorporado la inteligencia artificial a alguno de sus procesos, un punto por encima de la media europea, mientras que entre las pymes el porcentaje baja al 7%, según el Observatorio Nacional de Tecnología y Sociedad, adscrito a la entidad pública Red.es. Las compañías se van familizarando con la IA pero, según la experta Elena González-Blanco, en áreas como banca o seguros «apenas hay datos publicados en corpus abiertos que sirvan para entrenar a las máquinas» porque las empresas no suelen compartirlos. «Solo aquellas que han hecho entrenamiento de IA en sus propios sistemas han sido capaces de desarrollar estos algoritmos. Cada vez que se hace una implementación de este tipo de proyectos en una empresa es como si se empezase de cero». Queda margen de mejora.

después las empresas privadas llevarán a sus departamentos de I+D si consideran que puede ser beneficioso», expone.

En este sentido se pronuncia Nieves Ábalos, fundadora de Monoceros Labs, un estudio de innovación especializado en las experiencias conversacionales que trabaja en la creación de aplicaciones de voz para hispanohablantes para asistentes como Amazon Alexa y Google Assistant. «En España no hemos tenido grandes empresas que empujen desde lo privado y, desde lo público, la investigación es difícil. Hay pocos recursos para crear corpus de datos o meter músculo para que podamos llegar más lejos. El español es una de las lenguas más importantes del mundo y aquí somos siempre los últimos», lamenta.

Este estudio, puesto en marcha a finales de 2017, es responsable, entre otras, de aplicaciones como el juego infantil 'Veo Veo', adaptado para dispositivos de Amazon Alexa, o 'Historias para recordar', para Google Assistant, que ayuda a ejercitar la memoria a las personas mayores. «Entrenamos a los algoritmos en español y diseñamos toda la experiencia pensando en este idioma», detalla. Ábalos cree que es imprescindible crear soluciones en español y no solamente adaptadas al español, sobre todo en lo relativo a interfaces conversacionales. ¿La razón? «Si casi no vale una aplicación traducida del español de España al de México, menos del inglés al español. Para dar una solución satisfactoria a los usuarios y que puedan usar estas tecnologías en su día a día tenemos que hacer experiencias diseñadas desde el español», comenta. Cabe recordar que asistentes virtuales como Alexa o Google Assistant nacieron primero en inglés, en 2014 y 2016 respectivamente, y no se lanzaron en español hasta años después.

La partida por el dominio de esta tecnología se libra en un tablero en el que el Viejo Continente está desposicionado y nuestro país apenas ha utilizado los datos del español para entrenar a los algoritmos. «Europa se preocupa por las cuestiones éticas y de regulación, pero cuenta con pocas empresas grandes que sean tractoras en el mercado. Es verdad que se ha trabajado muchísimo en los temas de traducción y multilingüismo por razones obvias de comunicación entre los países, pero ahora se está quedando atrás», comienza por destacar González-Blanco. Ante esta situación, cada país ha optado por centrar el tiro en determinadas áreas. Alemania, por ejemplo, ha vuelto la mirada a las tecnologías de IA aplicadas al sector automovilístico por el despliegue del coche autónomo, mientras que, a juicio de la experta, España carece de un posicionamiento claro.

Fortalezas y déficits

Más futuro que presente

585

En todo el mundo más de 585 millones de personas hablan español, ya sea como lengua nativa, segunda o extranjera, lo que supone el 7,5% de la población mundial, según datos del Instituto Cervantes referidos a 2020

30%

El español todavía representa menos del 30% del mercado mundial de las tecnologías de procesamiento de lenguaje natural

500

El Gobierno ha destinado 500 millones de euros para el periodo 2021-2023 a la inteligencia artificial dentro del Plan de Recuperación, Transformación y Resiliencia. Esta cantidad representa el 0,7% de los 70.000 millones que recibirá España de Bruselas

34,5%

Los 'papers' científicos sobre procesamiento del lenguaje e inteligencia artificial aplicada al lenguaje han aumentado un 34,5%entre 2019 y 2020

Retos pendientes

Pero antes de pensar en cómo liderar el uso de la lengua española en la IA, nuestro país debe resolver problemas de base que lastran cualquier intento de progreso. El reto más acuciante es estrechar lazos entre las universidades y el sector empresarial. «Todo el conocimiento científico se tendría que materializar en emprendimiento, ayuda a las empresas, creación de puestos de trabajo, etc. Es esencial para retener talento técnico y para darle valor económico a los avances científicos porque en España hay grupos de investigación muy buenos en ese sentido», asegura González-Blanco. Otro de los obstáculos es la disponibilidad de datos. «Que no haya demasiados corpus públicos puede suponer un freno porque no se tiene el nutriente del que bebe la IA. El hecho de que hubiese corpus públicos y de fácil acceso sería un avance importante», defiende Moya.

Es también una cuestión de mentalidad. «Aquí cada uno tiene sus datos bajo llave y no hay una cultura de compartir. Los proyectos pequeños son muy caros y tienen resultados limitados, por lo que hay que montar grandes infraestructuras de IA donde los datos puedan abrirse y reutilizarse para tener una iniciativa de crecimiento que pueda empujar esta tracción. O se hace algo grande o no competiremos con corporaciones que pueden hacerlo muchísimo más rápido que nosotros», destaca.

El lado positivo, en su opinión, es que España ahora está en un buen momento para abordar estos desafíos porque cuenta con la Secretaría de Estado de Digitalización e Inteligencia Artificial, encabezada por Carme Artigas, y con el dinero procedente de Bruselas.

Dentro del Plan de Recuperación, Transformación y Resiliencia, el Gobierno ha dedicado el componente 16 a la Estrategia Nacional de Inteligencia Artificial, a la que destinará 500 mi-

▶▶▶

llones de euros para el periodo 2021-2023, el 0,7% de los 70.000 millones que recibirá de Bruselas. Entre sus objetivos está situar a España como país puntero en este ámbito y liderar a nivel mundial el uso de la lengua española en la IA. Propone para ello el Plan de Tecnologías de Lenguaje Natural, dotado con 28 millones de euros en tres años, que incluye la creación de un centro de IA en español para poner a disposición de las empresas recursos que les permitan hacer uso de estas tecnologías y la elaboración de convenios con contenido económico con las principales instituciones generadoras de datos y corpus.

Fuentes de la Secretaría de Estado de Digitalización e Inteligencia Artificial matizan que a estos 28 millones habría que sumar inversiones en materia de I+D+i contempladas en otros componentes más horizontes que también se destinarán al uso del español en la IA.

Explican que la creación de un corpus del español es esencial para el desarrollo de la industria digital en español. «El hecho de que los algoritmos se estén entrenando con corpus primordialmente en inglés puede introducir sesgos que no tengan en cuenta las especificidades de la comunidad hispanohablante», resaltan. El impulso del español en la IA permitirá una menor dependencia de multinacionales o entidades privadas extranjeras, lo que tendrá «relevantes implicaciones en materia de acceso, calidad y seguridad». Y, en lo que respecta a la transferencia de la investigación basada en el procesamiento de lenguaje natural del español, «permitirá impulsar aplicaciones y productos desarrollados para diversos ámbitos y sectores de utilidad para España y para todos los países hispano hablantes», indican.

Mercado amplio

Las oportunidades de negocio son inmensas porque la explotación de los datos del español abriría las puertas a un mercado potencial con más de 585 millones de hispanohablantes. «Es una economía de escala muy grande, ya que una solución para el español puede venderse en un amplio espectro de países», subraya Luis de la Fuente (Unir). Esta universidad lidera el proyecto Plentas, basado en diseñar un sistema de IA que evalúe automáticamente la respuesta de los alumnos a una pregunta hecha por el profesor. «En un caso ideal, el sistema se conecta con una plataforma educativa, el profesor pide escribir en dos frases cómo fue el descubrimiento de América, por ejemplo, y a los diez minutos, los alumnos reciben una respuesta con la nota y los motivos», detalla. Soluciones como esta –aún en fase de investigación básica– podrían comercializarse a un número elevado de centros educativos. «Si este despliegue lo llevásemos a América Latina, los alumnos que se podrían aprovechar sería brutal», subraya De la Fuente.

Habría beneficios a varios niveles. «Desde un punto de vista tecnológico nos posicionaríamos como líderes en un área de la IA que todavía está virgen, en la que podemos aportar va-

TRADUCCIÓN SIMULTÁNEA PARA ROMPER BARRERAS

El entrenamiento de los algoritmos en distintos idiomas más allá del inglés será un salto cualitativo para la mejora de las traducciones automáticas y simultáneas. Así lo defiende Ricardo Moya, experto IA de IMF: «Que la inteligencia artificial pueda romper las barreras lingüísticas y permitirnos mantener conversaciones con personas que no dominan nuestra lengua beneficiará no solo a la comunidad de hispanohablantes sino a todo el mundo».

MODELOS DE LENGUAJE

Supercomputación para captar todos los matices

L. MONTERO

Un adjetivo como «brutal» puede tener distinto significado en función del contexto en el que se utilice. No es lo mismo decir «la serie es brutal» que «el brutal asesinato». Los hablantes de carne y hueso somos capaces de percibir la diferencia, pero ¿también la inteligencia artificial? Por increíble que parezca, la respuesta es que sí. Los modelos del lenguaje reproducen el uso de la lengua y permiten conocer el significado real de las palabras, incluso de las frases enteras. Hasta ahora se han hecho modelos del lenguaje sobre todo del inglés, pero este desequilibrio está cerca de cambiar gracias al trabajo conjunto de dos instituciones españolas.

El Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC) y la Biblioteca Nacional de España (BNE) han unido fuerzas para generar un modelo del lenguaje del español y de otras lenguas del estado. Lo hacen por encargo de la extinta Secretaría de Estado para el Avance Digital, en el marco del Plan de Impulso de las Tecnologías del Lenguaje.

«La BNE es una fuente muy rica y valiosísima para hacer un modelo del lenguaje por la cantidad ingente de datos que tiene, mientras que el BSC cuenta con el supercomputador MareNostrum para el procesado masivo. Ya que tenemos los recursos, es cuestión de aprovecharlos», subraya Quim Moré, investigador del departamento de CASE del BSC y uno de los responsables del proyecto en el centro. El Archivo Web de la BNE es la colección formada por los sitios web con dominio .es, incluidos documentos, imágenes, vídeos, etc.

«Nosotros solo necesitábamos el texto. Desde la BNE aplicaron un proceso de extracción de los datos textuales, que ya se han transferido al supercomputador. Tiene almacenados un total de 45 terabytes», cuenta Moré. Completada esta fase, el siguiente paso es su procesado para generar el modelo del lenguaje a través de las tecnologías del procesamiento del lenguaje natural.

45 terabytes tiene almacenados MareNostrum

La idea es crear lo que se ha hecho con Google Bert, la actualización del algoritmo de este motor de búsqueda. «El modelo del lenguaje lo que hace es que, según el contexto en que aparece una frase, interpreta su sentido. Esto tiene una aplicación a la traducción automática», indica Moré. Pero las posibilidades no se agotan ahí. «Pongamos por caso que tenemos un asistente autónomo de coche, se encuentra con una información lingüística que se puede interpretar de una forma u otra y el coche

lor diferencial y, además, ser un puente con América Latina», dice Elena González-Blanco. Para millones de empresas con sede en España y en el continente americano supondría la posibilidad de automatizar procesos y aumentar su negocio. Y eso sin olvidar que España podría convertirse en un 'hub' de atracción de talento tecnológico. «Tenemos calidad de vida y condiciones que no solo atraen turismo, sino que pueden ser estratégicas para montar startups y atraer talento de investigación y tecnología que puede ser clave para que los proyectos crezcan y se vuelvan exponenciales», señala. El potencial del español como catalizador de la competitividad nacional en IA es incalculable.

MareNostrum, el supercomputador más potente de España

tiene que actuar según esta interpretación. El modelo del lenguaje ayuda a que tome la decisión correcta porque interpreta el sentido según el contexto», dice.

«Gran avance»

Esta iniciativa significa, en palabras de Quim Moré, la oportunidad de «tener un modelo del lenguaje del español tan potente como los que puede hacer Google», lo que supone «un gran avance en la aplicación de la IA en español, sobre todo basada en conocimiento de la realidad con el lenguaje». Las bases para un futuro mejor empiezan a construirse.

PROYECTO Un plan para preservar la unidad del idioma

LEIA, o cómo proteger nuestra lengua en el desembarco de las máquinas

La RAE, apoyada por grandes firmas tecnológicas, lidera una iniciativa que une la lengua y la inteligencia artificial para el cuidado del español

L. MONTERO

Chatbots, sistemas de mensajería instantánea, asistentes virtuales... las máquinas forman parte de nuestro día a día. Con el propósito de enseñarlas a hablar un correcto español nació en 2019 el proyecto Lengua Española e Inteligencia Artificial (LEIA), ideado y liderado por la Real Academia Española (RAE). «Tiene como fin principal cuidar el uso de un correcto español en los medios tecnológicos y así evitar que se pierda la unidad que permite que más de 580 millones de personas podamos comunicarnos en nuestra lengua sin dificultades. A día de hoy hay más millones de máquinas hablando español que hispanohablantes por lo que es importante unir la lengua y la inteligencia artificial para cuidar nuestro idioma», asegura Santiago Muñoz Machado, director de la RAE y presidente de la Asociación de Academias de la Lengua Española (ASALE). Además de crear e impulsar el proyecto, la RAE aporta su conocimiento profundo de la lengua y sus valiosos recursos, como los diccionarios y otras obras, corpus o recopilaciones inmensas de textos, y bases de datos con consultas lingüísticas resueltas.

Impacto

La primera compañía en sumarse a la iniciativa fue Telefónica. Ahora cuenta en sus filas con Microsoft, Amazon, Google, Twitter y Facebook. «Se está ya trabajando con los socios tecnológicos para mejorar el español de, por poner un ejemplo, sus aparatos y asistentes de voz», indica Muñoz Machado. Pero hay más acciones. «El proyecto cuenta con el apoyo del Gobierno, en primer lugar del de España. Para que los ciudadanos puedan disfrutar y tener a su disposición un español correcto en cualquier ámbito digital, la Academia está trabajando en crear un observatorio para detectar neologismos en la Red, herramientas que permitan conocer las distintas variedades y usos del español en internet, y aportar una mayor accesibilidad a los usuarios», subraya Muñoz Machado.

Con procesos relacionados con la IA como el aprendizaje de máquina, se pueden crear también, en palabras de Muñoz Machado, «herramientas de análisis sintáctico y léxico que permitan detectar automáticamente el leísmo o el laísmo, o errores de concordancia. También identificar y extraer por sí solas derivados, neologismos o extranjerismos. Todo ello ayuda a mejorar el funcionamiento de correctores y facilita la elaboración de obras lingüísticas». Además, la RAE espera que LEIA ayude a que el español alcance un posicionamiento similar al del inglés en IA.

«Tenemos que ir al 'español desde el diseño'», dice Chema Alonso, director de Consumo Digital de Telefónica y director Técnico de LEIA. En su opinión, el número de hablantes que tiene y el que sea la lengua oficial en más de 20 países «garantiza, hasta cierto punto, que nunca se va a ignorar», pero cree que «hay riesgo de que cada empresa tecnológica fomente un español distinto impulsado por intereses comerciales y no por fines culturales y sociales, o simplemente por el modo en que la tecnología trata nuestra lengua».

Sostiene que un alto porcentaje de personas escriben documentos en español a través de herramientas de ofimática de las grandes tecnológicas. «En algunos casos usan diccionarios y léxicos que no reconocen alrededor del 10% de las palabras que sí están reconocidas por la RAE», detalla. Y eso tiene sus consecuencias. «Dejaremos de usar estas palabras correctas y el español irá empobreciéndose progresivamente. Y este riesgo se incrementa mucho más si tenemos en cuenta los cientos de millones de personas que escriben mensajes en aplicaciones de mensajería y redes sociales que tampoco usan los recursos de la RAE», dice. Telefónica está integrando los recursos lingüísticos de la RAE en AURA, su asistente virtual basado en IA, para que todo lo que diga o escriba sea de acuerdo con la RAE. «Como responsable de la dirección técnica de LEIA, Telefónica está coordinando con las grandes tecnológicas cómo pueden incluir los recursos lingüísticos de la RAE en sus productos y servicios», añade.

Garantía de calidad

Se está desarrollando un sello digital para que la RAE certifique que las herramientas tecnológicas y de IA de empresas, fundaciones y otros colectivos usan el español de manera apropiada