Revista de Comunicación de la SEECI (2026).
ISSN: 2695-5156

Recibido: 20/10/2026 --- **Aceptado:** 05/11/2026 --- **Publicado:** 16/11/2026

AUTOMATIZACIÓN LINGÜÍSTICA, SOBERANÍA TECNOLÓGICA Y JUSTICIA COGNITIVA: UNA CRÍTICA DE LAS INFRAESTRUCTURAS DEL LENGUAJE

LINGUISTIC AUTOMATION, TECHNOLOGICAL SOVEREIGNTY AND COGNITIVE JUSTICE: A CRITIQUE OF LANGUAGE INFRASTRUCTURES

Manuel De-la-Cruz-Recio: Universidad de Salamanca. España.
manueldelacruzrecio@usal.es

Cómo citar el artículo:

De-la-Cruz-Recio, Manuel. (2026). Automatización lingüística, soberanía tecnológica y justicia cognitiva: Una crítica de las infraestructuras del lenguaje. *Revista de Comunicación de la SEECI*, 59, 1-15. https://doi.org/10.15198/seeci.2026.59.e990

RESUMEN

Introducción: La automatización lingüística modifica las tareas comunicativas y los procesos de selección y circulación de formas de lengua. El artículo examina qué implica esta mediación para la soberanía tecnológica en el espacio iberófono. Objetivos: Precisar las condiciones bajo las cuales el control de las infraestructuras del lenguaje puede vincularse con justicia cognitiva. **Metodología:** Se desarrolla una investigación teórico-documental que articula teoría crítica de la tecnología, filosofía de la mediación y pensamiento caribeño. La lectura de las fuentes se organiza mediante categorías de representación, diseño, agencia y legitimidad epistemológica. Se examinan el filtrado de textos en C4, derivado de Common Crawl, y la especificación de portugués brasileño y europeo como variedades de destino en traducción automática. **Resultados:** El análisis distingue los efectos representacionales del filtrado y las posibilidades de especificar variedades de destino. Su interpretación, articulada con las premisas del marco filosófico, fundamenta las condiciones de pluralismo representacional, democratización del diseño y responsabilidad epistemológica. **Discusión:** El pensamiento caribeño permite interrogar la universalización de criterios históricos de reconocimiento. **Conclusiones:** La soberanía tecnológica requiere capacidad de intervención sobre las decisiones que organizan corpus, estándares y mediaciones. La justicia cognitiva orienta la evaluación

de esas decisiones. La aplicación de las condiciones propuestas requiere especificación y contraste en infraestructuras concretas.

Palabras clave: automatización lingüística; soberanía tecnológica; justicia cognitiva; traducción automática; infraestructuras digitales; pensamiento caribeño; pluralismo lingüístico.

ABSTRACT

Introduction: Linguistic automation reshapes communicative tasks and the selection and circulation of linguistic forms. This article examines what such mediation implies for technological sovereignty in Spanish- and Portuguese-speaking contexts. Objectives: To specify the conditions under which control over language infrastructures can be connected with cognitive justice. **Methodology:** A theoretical study based on documentary analysis and deductive reasoning brings together critical theory of technology, philosophy of technological mediation and Caribbean thought. Sources are examined through the categories of representation, design, agency and epistemological legitimacy. The study considers text filtering in C4, derived from Common Crawl, and the specification of Brazilian and European Portuguese as target varieties in machine translation. **Results:** The analysis distinguishes the representational effects of filtering and the possibilities for specifying target varieties. Interpreted in conjunction with the premises of the philosophical framework, these findings inform the justification of representational pluralism, democratisation of design and epistemological responsibility. **Discussion:** Caribbean thought enables an examination of how historically situated criteria of recognition acquire universal authority. **Conclusions:** Technological sovereignty requires the capacity to intervene in decisions that organise corpora, standards and mediation. Cognitive justice guides the evaluation of these decisions. Applying the proposed conditions requires further specification and investigation in particular infrastructures.

Keywords: linguistic automation; technological sovereignty; cognitive justice; machine translation; digital infrastructures; Caribbean thought; linguistic pluralism.

1. INTRODUCCIÓN

La automatización lingüística interviene en la selección, normalización y circulación de formas de lengua a través de la traducción automática, los grandes modelos lingüísticos y las plataformas multilingües. El problema que aborda este artículo es cómo esa intervención afecta a las condiciones en las que las comunidades pueden producir y compartir conocimiento. La expansión de la inteligencia artificial generativa ha vuelto más visible este problema. Las advertencias de Bender *et al*. (2021, pp. 610-613) sobre la escala de los modelos y sus datos de entrenamiento permiten situarlo en las decisiones que sostienen la mediación automatizada. Entre esas decisiones, la constitución de los datos ocupa un lugar central. Floridi (2019, pp. 1-3) relaciona su naturaleza con los tipos de problemas que los sistemas pueden abordar, y Crawford (2021, pp. 95-98) los sitúa dentro de procesos materiales y sociales de extracción y clasificación. En el ámbito lingüístico, atender a su procedencia permite preguntar qué

usos se incorporan a los sistemas y cuáles permanecen fuera. Por ello, una evaluación de la automatización debe considerar cómo se construyen sus recursos lingüísticos.

La representación de la diversidad depende de cómo los corpus recogen variedades regionales, registros y géneros discursivos. En su revisión del procesamiento del lenguaje natural, Zampieri *et al*. (2020) muestran la necesidad de distinguir lenguas, variedades y tareas de evaluación. Una herramienta disponible en español o portugués requiere, por ello, un examen de los usos que reconoce y de aquellos que toma como referencia. Esta distinción desplaza la atención desde la presencia de una lengua hacia las condiciones de su representación. Cuando se traslada al ámbito lingüístico, la soberanía tecnológica exige atender a esas condiciones. Disponer de infraestructuras propias, capacidad computacional o autonomía frente a proveedores externos puede ampliar el margen de actuación de una institución. El ejercicio de esa capacidad incluye decisiones sobre qué corpus se utiliza y qué criterios validan sus resultados. La pregunta por la soberanía alcanza así al diseño, a los estándares y a la autoridad para definirlos, en la medida en que las decisiones técnicas incorporan intereses y prioridades sociales (Feenberg, 1999, pp. 88-90).

Esta pregunta se sitúa aquí en el espacio iberófono. El término designa operativamente los entornos comunicativos articulados por el español y el portugués, considerados en su diversidad interna y en sus relaciones con otras lenguas. Las lenguas indígenas, africanas y criollas mantienen su especificidad dentro de esas relaciones. Se trata de un ámbito heterogéneo, atravesado por experiencias históricas y posiciones políticas distintas. Esta delimitación permite examinar cómo la autonomía tecnológica de una lengua de amplia difusión puede coexistir con desigualdades entre sus hablantes y con la subordinación de otras comunidades lingüísticas.

Para interrogar esas desigualdades se recurre al pensamiento caribeño como perspectiva crítica sobre el reconocimiento. La sobrerrepresentación de una figura particular de lo humano, examinada por Wynter (2003), ofrece un punto de partida para comprender cómo ciertos criterios adquieren autoridad más allá de las condiciones históricas en las que se formaron. Su aplicación a las infraestructuras del lenguaje es una interpretación desarrollada en este artículo, que permite preguntar qué concepciones de hablante, competencia y conocimiento se toman como referencia. La función del Caribe en el argumento queda delimitada por esa aportación filosófica.

La justicia cognitiva permite formular el criterio normativo de esta pregunta. Visvanathan (2009) la vincula con el reconocimiento de formas plurales de conocimiento y de las prácticas de vida que las sostienen. En este artículo se emplea para examinar si las infraestructuras lingüísticas permiten que esas formas de conocimiento sean expresadas y discutidas en condiciones de reconocimiento. Conviene distinguir aquí la pluralidad de repertorios lingüísticos y la de conocimientos. Representar una variedad puede favorecer la expresión de una experiencia; la interpretación y la valoración de esa experiencia requieren, además, examinar los criterios de legitimidad que se le aplican.

El problema de investigación se encuentra, así, en la articulación entre diseño técnico, mediación y legitimidad del conocimiento. La teoría crítica de la tecnología ofrece

recursos para examinar quién configura los sistemas, mientras que la filosofía de la mediación permite estudiar cómo esas configuraciones intervienen en la experiencia. El pensamiento caribeño introduce la pregunta por la autoridad histórica de los criterios con los que se reconoce a sujetos y saberes. El trabajo pone en relación estas perspectivas atendiendo a sus aportaciones específicas y a las cuestiones que comparten. La contribución principal consiste en desplazar la discusión sobre soberanía tecnológica desde la posesión de capacidades infraestructurales hacia las condiciones de legitimidad de su ejercicio, articuladas mediante pluralismo representacional, democratización del diseño y responsabilidad epistemológica. Para fundamentarlas, el artículo combina reconstrucción teórica y dos microcasos documentales que concretan decisiones sobre datos y resultados. Se examinan el filtrado de textos en C4, derivado de Common Crawl, y la especificación de portugués brasileño y europeo como variedades de destino en traducción automática. El alcance empírico de cada caso queda circunscrito a los materiales y sistemas estudiados por sus autores.

2. OBJETIVOS

La pregunta que articula el trabajo es qué condiciones normativas debe satisfacer una concepción de soberanía tecnológica orientada a evitar que la automatización lingüística reproduzca exclusión epistemológica y jerarquización lingüística en el espacio iberófono. Se examina, en particular, la capacidad de discutir las decisiones mediante las que se ejerce el control de la infraestructura y las consecuencias que tienen para las comunidades afectadas. El objetivo general consiste en elaborar un marco analítico para estudiar la relación entre automatización lingüística, soberanía tecnológica y justicia cognitiva. El recorrido distingue cómo se representa la diversidad en los corpus, cómo se establecen criterios de adecuación en los resultados y cómo pueden justificarse las decisiones que afectan a ambas dimensiones. La base documental del análisis está constituida por investigaciones publicadas; las evaluaciones de rendimiento comentadas corresponden a esos estudios.

3. METODOLOGÍA

El trabajo adopta un diseño de investigación teórico-documental con razonamiento deductivo. Examina la relación entre automatización lingüística, soberanía tecnológica y justicia cognitiva mediante la articulación de fuentes filosóficas y resultados de investigaciones publicadas. La unidad de análisis son las decisiones mediante las que corpus, estándares y plataformas organizan condiciones de comunicación. A partir de las premisas del marco teórico y de las operaciones documentadas en los microcasos, se justifican exigencias sobre representación, intervención y responsabilidad. La argumentación se evalúa por la precisión de sus distinciones, la adecuación de las fuentes al problema y la explicitación del paso de las premisas a las condiciones normativas propuestas.

El corpus teórico reúne textos sobre diseño técnico, mediación y reconocimiento. La lectura de Feenberg (1999, pp. 88-90, 145) se concentra en la inscripción de intereses en el diseño y en sus posibilidades de democratización. La intervención de los artefactos en la percepción y la acción se examina a partir de Verbeek (2008, pp. 389-390). Los textos de Fanon, Glissant y Wynter aportan categorías para estudiar la

dimensión histórica de los criterios de reconocimiento. Estas funciones delimitan los pasajes relevantes de cada obra y su lugar en el argumento.

La selección bibliográfica responde a una estrategia orientada por el problema. Se consideran pertinentes los trabajos que permiten examinar una decisión de diseño, describir una práctica de mediación o justificar un criterio de reconocimiento. La literatura técnica y documental utilizada se concentra en trabajos de 2018-2024, con Papineni *et al*. (2002) como referencia para explicar BLEU. Las fuentes filosóficas se seleccionan por las categorías que aportan y por su relación con la pregunta normativa. El corpus tiene carácter selectivo y sirve para reconstruir los argumentos empleados; una revisión sistemática de las posiciones existentes queda fuera de su alcance. Las fuentes sobre automatización lingüística aportan descripciones de prácticas y decisiones técnicas. Baack (2024) permite examinar la producción de datos web a través de Common Crawl, mientras que Dodge *et al*. (2021) analizan el efecto del filtrado en C4. Las dimensiones de documentación de conjuntos de datos se apoyan en Bender y Friedman (2018) y Gebru *et al*. (2021). Lakew *et al*. (2018) y Zampieri *et al*. (2020) sitúan el problema de la variación lingüística, y Nee *et al*. (2022) proporcionan criterios para discutir sus consecuencias para los usuarios desde la justicia lingüística.

La lectura se organiza mediante las categorías de representación, diseño, agencia y legitimidad epistemológica. Con ellas se examinan los usos y comunidades que aparecen en los datos, la autoridad para establecer estándares y criterios de evaluación, las posibilidades de intervención de los sujetos y los conocimientos que se reconocen como pertinentes. Estas categorías orientan la interpretación de las fuentes y la comparación entre decisiones. Su función es hacer explícitas las preguntas aplicadas a los materiales en un análisis cualitativo de argumentos y resultados publicados.

Los dos microcasos se seleccionan para examinar la conservación de textos durante la preparación de un corpus y la especificación de una variedad al producir una traducción. El primero reconstruye el contraste entre versiones de C4 con y sin una lista de bloqueo, estudiado por Dodge *et al*. (2021). El segundo examina la comparación de Lakew *et al*. (2018) entre sistemas que reúnen datos de portugués brasileño y europeo sin distinguir las variedades y sistemas que incorporan una etiqueta de destino. La comparación entre los casos relaciona decisiones de selección y especificación procedentes de investigaciones independientes. El estudio de C4 aporta evidencia sobre un corpus en inglés; el experimento de traducción utiliza otros materiales y se refiere a variedades del portugués. Las magnitudes obtenidas conservan el significado que tienen en cada estudio. Su aplicación al espacio iberófono consiste en identificar decisiones susceptibles de examen en otros contextos, cuyo funcionamiento exige una investigación situada.

La derivación de las condiciones normativas parte de la premisa de que las comunidades afectadas deben poder participar en la discusión de los criterios que representan sus prácticas y conocimientos. Puesto que las decisiones de selección y normalización intervienen en esa representación, su justificación debe someterse a un examen en el que puedan participar dichas comunidades. De esta relación se derivan

exigencias relativas a la composición de los datos, a la autoridad sobre el diseño y a la responsabilidad por los criterios empleados. Los resultados publicados permiten localizar las decisiones a las que se aplican esas exigencias y discutir su alcance. La valoración de su justicia incorpora la premisa normativa explícita, junto con la interpretación de sus efectos. Las conclusiones constituyen una propuesta analítica cuya aplicación podrá contrastarse mediante estudios sobre corpus, plataformas, traducción automática, modelos lingüísticos o gobernanza comunitaria de datos.

Antes de continuar, se quiere manifestar que el manuscrito fue concebido, argumentado y redactado por el autor. Tras la comunicación editorial preliminar, se empleó ChatGPT (OpenAI, GPT-5.5 Pro; junio de 2026) exclusivamente como herramienta de apoyo editorial para la reorganización estructural del texto, la adaptación a la plantilla de la revista, la detección de inconsistencias formales, la revisión de coherencia interna y la mejora de la redacción académica. La herramienta no se utilizó para generar datos, producir resultados empíricos, sustituir el análisis conceptual del autor, inventar referencias ni formular conclusiones autónomas. Todas las sugerencias fueron revisadas, contrastadas con las fuentes citadas y aceptadas o descartadas por el autor, quien asume la responsabilidad íntegra del contenido final.

4. RESULTADOS

Los resultados desarrollan el recorrido deductivo del argumento. El análisis de la infraestructura localiza decisiones sobre el lenguaje, los microcasos reconstruyen dos operaciones técnicas y la última subsección justifica las condiciones normativas aplicables a su examen desde las premisas de reconocimiento y participación.

4.1. Automatización lingüística como infraestructura sociotécnica

El análisis de la automatización lingüística como infraestructura sociotécnica relaciona sus componentes técnicos con las prácticas e instituciones que les atribuyen funciones. Esta lectura, apoyada en Feenberg (1999, pp. vii-viii), permite examinar las condiciones que una tecnología establece para otras actividades junto con las tareas que realiza. La traducción automática puede facilitar tareas comunicativas e intervenir en qué formas de lenguaje resultan reconocibles, normalizables y circulables. La mediación descrita por Verbeek (2008, pp. 389-390) permite interpretar esa intervención como participación de los artefactos en las prácticas humanas. Desde la perspectiva adoptada aquí, la automatización contribuye a estabilizar condiciones de valor lingüístico dentro de arquitecturas técnicas. El procesamiento textual queda así vinculado con una política de la legibilidad, entendida como organización de lo que los sistemas hacen reconocible y disponible.

Cuando entra en la economía digital, el lenguaje opera como recurso computacional, materia prima de entrenamiento y soporte de interoperabilidad. Crawford (2021, pp. 95-98) muestra que los datos de entrenamiento se construyen mediante operaciones de extracción y clasificación. En el caso del lenguaje, esas operaciones seleccionan textos, registros y fuentes para fines determinados. La infraestructura trabaja con materiales delimitados, cuya composición condiciona qué usos están disponibles para el sistema. La delimitación de esos materiales constituye una decisión de diseño, incluso cuando se presenta como una operación rutinaria. Los códigos técnicos

estabilizan intereses y prioridades sociales que pueden llegar a parecer evidentes (Feenberg, 1999, pp. 88-90). Trasladada a la automatización lingüística, esta perspectiva invita a examinar qué se considera ruido, qué merece conservarse y qué se utiliza como referencia. La justificación de cada criterio debe atender a sus fines y a las alternativas que deja fuera.

Los efectos del diseño se desarrollan en las prácticas de uso. Verbeek (2008, pp. 389-390) analiza cómo los artefactos participan en la configuración de la percepción y la acción. Esta perspectiva permite examinar cómo encuentran y utilizan los sujetos las posibilidades que la infraestructura les ofrece. En la automatización lingüística, interesa conocer de qué manera una salida sugerida, una corrección o un criterio de evaluación orienta una práctica comunicativa. Si una propuesta automática se convierte en referencia por defecto, el margen de intervención depende de que pueda examinarse, rechazarse o modificarse. La redistribución de agencia designa aquí un cambio en esas condiciones de decisión. Este supuesto analítico conserva la capacidad de los usuarios para producir sentido y sitúa la relación entre automatización y poder en las posibilidades concretas de intervenir sobre una propuesta. Permite, por ello, examinar el trabajo profesional atendiendo a la autoridad atribuida a los resultados automáticos.

El análisis de la infraestructura conduce a examinar cómo se seleccionan los materiales que alimentan el sistema y cómo se orientan las prácticas mediante sus resultados. La primera pregunta se concreta en la regla de filtrado aplicada al construir C4; la segunda, en la posibilidad de especificar una variedad portuguesa de destino. Los apartados siguientes reconstruyen un estudio de cada operación para discutir sus implicaciones y sus límites.

4.2. De Common Crawl a C4: el filtrado como decisión de representación

El primer microcaso examina el filtrado de documentos en el paso del archivo de Common Crawl al *Colossal Clean Crawled Corpus* (C4), un conjunto de datos preparado para entrenar modelos lingüísticos. Baack (2024, seccs. 4 y 6) distingue la recopilación de páginas web de las decisiones posteriores de quienes construyen corpus de entrenamiento. Cada etapa introduce criterios que delimitan la composición del recurso resultante. La reconstrucción de Baack (2024, seccs. 3 y 6.1), apoyada en entrevistas al director y al responsable del rastreo realizadas en 2023 y en documentación pública, describe un procedimiento que prioriza dominios según su centralidad armónica. Su posición en la red de enlaces influye en la probabilidad de incluirlos y en cuántas páginas se recogen de ellos. Esta prioridad responde, entre otras razones, al propósito de reducir el *spam*. La evaluación de la representatividad sociolingüística requiere examinar además la composición de los materiales recopilados y los filtros utilizados en su preparación posterior.

El efecto de uno de esos filtros se examina en el estudio de Dodge *et al*. (2021, pp. 1287, 1291-1292). C4 procede de la captura de Common Crawl de abril de 2019. Los investigadores comparan C4.EN con C4.EN.NOBLOCKLIST, que mantiene los demás filtros y omite la exclusión de documentos que contienen términos de una lista de bloqueo de vocabulario considerado ofensivo. El contraste permite aislar el efecto de esa regla. Para analizar su distribución, los autores asignaron categorías dialectales a

los documentos mediante un modelo probabilístico. Las tasas de eliminación comunicadas fueron del 42 % en *African American English* y del 6,2 % en *White-aligned English* (Dodge *et al*., 2021, p. 1292). Estas categorías corresponden a estimaciones sobre los textos; la identidad de sus autores queda fuera de esa medición. El resultado documenta una selección desigual dentro del corpus en inglés examinado.

La interpretación propuesta se concentra en la justificación del criterio de exclusión. Si se pretende eliminar contenido dañino, debe examinarse si la regla distingue ese contenido de usos lingüísticos pertinentes para la población destinataria. La desigualdad observada exige valorar las consecuencias de la regla y las alternativas disponibles. Cada exclusión debe poder justificarse atendiendo a la finalidad del corpus y a sus efectos sobre la representación. Para ello es necesario documentar tanto el procedimiento como aquello que deja fuera. Los *data statements* de Bender y Friedman (2018, pp. 587-590) vinculan la interpretación de los datos con sus características lingüísticas y con las poblaciones implicadas. Aplicado aquí, ese enfoque permite preguntar si un criterio de limpieza responde al uso previsto o si necesita ser revisado. La documentación hace examinable la decisión y proporciona una base para discutir si la selección resulta adecuada.

Estas tasas corresponden al corpus en inglés estudiado. Su utilidad para el marco iberófono consiste en precisar qué habría que investigar en cada corpus, atendiendo a la decisión que modifica su composición, a sus efectos y a su justificación. La pregunta desplaza la soberanía desde el acceso al recurso hacia la capacidad de revisar su preparación. El apartado siguiente aborda la especificación de la variedad que debe producir una traducción.

4.3. Traducción al portugués brasileño y europeo: la variedad de destino

El segundo microcaso examina cómo se especifica la variedad de destino. Lakew *et al*. (2018, pp. 158-160) comparan sistemas de traducción del inglés al portugués brasileño y europeo. Se considera aquí la condición experimental de recursos reducidos, en la que los sistemas se entrenan con textos paralelos de charlas TED procedentes de WIT3. La comparación enfrenta *Gen*, que reúne los datos de ambas variedades sin distinguirlas, con *Mul*, que utiliza conjuntamente esos datos añadiendo una etiqueta de destino al entrenar y traducir. Ambos se entrenan con los datos reunidos; el segundo permite solicitar la variedad. En la tabla 4 del estudio, *Gen* obtiene 36,52 puntos BLEU para portugués brasileño y 33,75 para europeo; *Mul* alcanza 37,86 y 38,42, respectivamente. Lakew *et al*. (2018, p. 160) identifican ambas mejoras como significativas. Distinguir las variedades mejora aquí la puntuación y permite mantener sus datos de entrenamiento en un mismo conjunto.

La interpretación de esa mejora depende de qué mide BLEU. La métrica compara secuencias de palabras de las traducciones producidas con traducciones de referencia (Papineni *et al*., 2002, pp. 311-313). El incremento expresa un mejor resultado en esa comparación. La valoración de la adecuación comunicativa requiere justificar, además, la elección de las referencias y su relación con los destinatarios. La exigencia de coherencia con una variedad debe responder a la finalidad de la traducción. Cuando

un encargo requiere una variedad determinada, interesa que el sistema permita solicitarla; el uso de repertorios mixtos exige, por su parte, evaluar si las categorías disponibles se ajustan a esa práctica. La mezcla de formas debe juzgarse respecto de la situación comunicativa. La cuestión analítica es quién establece esa finalidad y cómo puede revisarla, de modo que la evaluación técnica se relacione con las necesidades de uso.

La capacidad de intervención sobre una herramienta que traduzca al portugués comprende la definición de las opciones de destino y la selección de los materiales de referencia. Una política de soberanía tendría que considerar qué usuarios pueden discutir esas decisiones y qué autoridad tienen para modificarlas. Hacer explícitos los fines comunicativos conecta la disponibilidad tecnológica con las decisiones sobre su configuración. El experimento analizado ofrece una alternativa de diseño documentada para portugués brasileño y europeo. Su extensión a otras variedades del portugués, al español o a otras arquitecturas requiere comprobaciones específicas. El caso de C4 examinaba una regla que decide qué textos se conservan; este identifica una decisión sobre qué variedad se solicita y con qué referencias se evalúa. La justificación de quién debe intervenir en ambas decisiones conduce a la cuestión del reconocimiento.

4.4. Del diagnóstico a la propuesta normativa: tres condiciones de soberanía tecnológica

La cuestión del reconocimiento permite precisar el alcance político de esas decisiones. En la lectura de Fanon (1952/2009, cap. I), hablar implica entrar en un mundo socialmente jerarquizado. Esta perspectiva permite preguntar cuándo un juicio sobre una forma lingüística afecta también al valor atribuido a la persona y a la experiencia que expresa. Wynter (2003) desplaza esa crítica hacia la sobrerrepresentación de una forma histórica de humanidad como si fuera universal. Desde esta perspectiva, el pensamiento caribeño ayuda a interrogar las ideas de sujeto, comunidad, lengua y conocimiento que operan en una infraestructura lingüística. La aplicación al problema tecnológico examina la autoridad que adquieren determinados criterios de reconocimiento. Este es un nivel de análisis distinto del recuento de una variedad en un corpus. La advertencia de Gordon (2006, pp. 4-5) sobre la conversión de los límites de una disciplina en medida de las demás permite precisar ese riesgo de universalización.

Las categorías de relación y opacidad de Glissant (1997, pp. 189-195) introducen un límite a esa pretensión de universalidad. El reconocimiento admite que la experiencia del otro exceda las categorías con las que se la interpreta. Aplicado al problema del artículo, este planteamiento exige distinguir la explicación de las decisiones técnicas y el alcance de las clasificaciones utilizadas. Los responsables de una infraestructura deben hacer examinables sus criterios y reconocer aquello que las categorías empleadas dejan sin representar.

La justicia cognitiva proporciona un criterio para evaluar esa reducción. Visvanathan (2009) vincula el reconocimiento de los conocimientos con las formas de vida que los sostienen. Desde esta premisa, las comunidades deben participar como interlocutoras sobre el sentido de los materiales que aportan. Este compromiso de reconocimiento

fundamenta las condiciones normativas propuestas. Los microcasos localizan las decisiones a las que se aplican, y su justificación se desarrolla mediante la relación entre esas decisiones y la premisa de participación. La delimitación de la justicia cognitiva requiere precisar su relación con nociones próximas. Nee *et al*. (2022) orientan la justicia lingüística hacia el acceso y la participación con distintos repertorios. La justicia epistémica atiende a la atribución injusta de credibilidad y a las desventajas de comprensión ligadas a los recursos interpretativos compartidos, que afectan a las personas como sujetos de conocimiento (Fricker, 2007, caps. I y VII). La justicia cognitiva sitúa el acento en la legitimidad de formas plurales de conocimiento. Estas perspectivas se relacionan por su atención al reconocimiento, preservando sus diferentes objetos y niveles de análisis.

La soberanía de datos plantea, en el sentido operativo empleado aquí, quién tiene autoridad para decidir sobre la recopilación y utilización de los materiales. La justicia cognitiva añade el examen de las categorías con las que se los representa. Una comunidad puede intervenir en la gestión de sus datos y encontrar que sus conocimientos se interpretan mediante categorías ajenas a su experiencia. Esta posibilidad permite explicar por qué la soberanía tecnológica necesita atender conjuntamente a la autoridad sobre la infraestructura y a la justificación de las decisiones. Las tres condiciones siguientes concretan esa exigencia.

1. La primera condición, el pluralismo representacional, exige examinar qué lenguas, variedades, registros y comunidades aparecen en los corpus y bajo qué condiciones. El caso de C4 muestra que una regla de filtrado puede producir tasas de exclusión distintas entre categorías dialectales estimadas (Dodge *et al*., 2021, p. 1292). La exigencia normativa consiste en justificar la composición del conjunto respecto de su finalidad y de las comunidades afectadas. La evaluación debe relacionar la presencia de los usos con las decisiones que explican su inclusión o exclusión.

2. La democratización del diseño concierne a la autoridad para establecer las categorías y modificarlas. El experimento de traducción documenta una posibilidad de especificación técnica; la intervención de los destinatarios en su definición constituye una cuestión que debe examinarse por separado. Feenberg (1999, p. 145) sitúa la legitimación participativa entre las exigencias de las decisiones técnicas. Retomada aquí, esa exigencia implica que comunidades, profesionales, instituciones y responsables tecnológicos puedan discutir la configuración de las infraestructuras que intervienen en el reconocimiento y la circulación del conocimiento. La democratización debe alcanzar la documentación, la evaluación, la gobernanza y la definición de estándares.

3. La responsabilidad epistemológica obliga a examinar qué formas de conocimiento se vuelven reconocibles dentro de los sistemas y cuáles quedan excluidas o subordinadas. Su ejercicio requiere identificar los supuestos de legitimidad que organizan corpus, modelos y estándares. La documentación de un conjunto de datos facilita ese examen al hacer visibles aspectos de su composición y sus usos (Gebru *et al*., 2021, pp. 86-88). La condición propuesta añade la exigencia de responder por los criterios de selección, filtrado y normalización y por las jerarquías históricas de conocimiento que puedan reproducir.

Las tres condiciones se necesitan mutuamente porque responden a dimensiones distintas de una misma capacidad de intervención. La ampliación de un corpus puede coexistir con una autoridad concentrada; un proceso participativo puede conservar criterios de reconocimiento restrictivos; y la atribución de responsabilidad requiere facultades efectivas para revisar la infraestructura. Su articulación permite examinar tanto las decisiones como las condiciones en que pueden ser impugnadas. En el espacio iberófono, esa articulación exige atender a las jerarquías entre variedades metropolitanas y periféricas, lenguas mayoritarias y minoritarias, registros institucionales y formas populares o comunitarias de producción lingüística. Una política centrada en competir con el inglés o con plataformas globales podría mantener tales jerarquías. La Recomendación de la UNESCO (2022) sobre ética de la inteligencia artificial y la presentación de un informe sobre soberanía indígena de datos (UNESCO, 2023) aportan referencias sobre participación comunitaria y derechos lingüísticos y culturales. Desde ese registro, permiten situar la discusión de la gobernanza en las relaciones entre capacidades tecnológicas y comunidades afectadas.

El recorrido desemboca en una propuesta de soberanía tecnológica como política de las infraestructuras del lenguaje. Su objeto comprende corpus, estándares, plataformas y modelos, junto con los criterios que determinan qué formas de lengua y conocimiento circulan con autoridad. La tabla 1 concreta las tres condiciones en preguntas, objetos de análisis y evidencias posibles para su evaluación posterior.

Tabla 1.

Condiciones de soberanía tecnológica: preguntas y evidencias para el análisis

Dimensión	Pregunta de evaluación	Objeto técnico	Riesgo	Actores implicados	Evidencias posibles
Pluralismo representacional	¿Qué usos y comunidades se incluyen o excluyen, y con qué justificación?	Corpus y reglas de selección y filtrado.	Cobertura agregada que oculta exclusiones.	Responsables de los datos y comunidades lingüísticas.	Procedencia; composición antes y después del filtrado; justificación de exclusiones.
Democratización del diseño	¿Quién establece y puede revisar las opciones de variedad y los criterios?	Etiquetas, estándares, referencias y métricas.	Consulta sin capacidad efectiva de decisión.	Desarrolladores, instituciones, profesionales y representantes de usuarios.	Acuerdos de participación; facultades de revisión; cambios incorporados tras la consulta.
Responsabilidad epistemológica	¿Cómo se explican y rectifican las clasificaciones que afectan al conocimiento?	Anotación, validación y criterios de uso.	Descontextualización o legitimación de jerarquías.	Responsables de anotación, evaluación y uso, junto con las comunidades afectadas.	Registro de criterios; respuestas a impugnaciones; rectificaciones y seguimiento de sus efectos.

Fuente: Elaboración propia a partir del marco desarrollado en 4.4. Las evidencias y los actores se proponen para evaluaciones futuras. Su aplicación como instrumento de evaluación requiere validación empírica.

5. CONCLUSIONES Y DISCUSIÓN

El análisis permite concluir que una concepción de soberanía tecnológica orientada a evitar la exclusión epistemológica y la jerarquización lingüística debe considerar cómo se representan los usos, quién puede intervenir en el diseño y cómo se justifican los criterios de legitimidad. Aquí se propone articular estas exigencias mediante pluralismo representacional, democratización del diseño y responsabilidad epistemológica. La contribución precisa las dimensiones que una apelación a la soberanía debe someter a discusión cuando se aplica al lenguaje. Los dos casos expuestos concretan decisiones diferentes. Dodge *et al*. (2021, p. 1292) documentan un efecto desigual de una regla de exclusión entre categorías dialectales estimadas; la comparación de Lakew *et al*. (2018, p. 160) muestra que distinguir variedades de destino puede mejorar la puntuación frente a las referencias utilizadas. Los resultados conservan el alcance de cada investigación y delimitan decisiones cuya legitimidad requiere una evaluación adicional. Las condiciones normativas se justifican a partir de la interpretación de esas decisiones y de las premisas de reconocimiento y participación.

Esta contribución desplaza la discusión sobre soberanía tecnológica desde la capacidad infraestructural hacia la justicia cognitiva y mantiene una tensión entre decisiones técnicas delimitadas y crítica histórica del reconocimiento. La filosofía de la mediación atiende a las relaciones situadas entre sujetos y artefactos; una lectura caribeña interroga la autoridad de las categorías que organizan esas relaciones. Su puesta en diálogo permite precisar qué debe investigarse para conectar una decisión técnica con un orden de reconocimiento. La crítica del colonialismo de datos sitúa esa pregunta en una discusión más amplia sobre apropiación y poder. Couldry y Mejias (2019a, 2019b) examinan la extracción de valor a partir de la vida social, y Birhane y Guest (2021) relacionan la descolonización de las ciencias computacionales con una revisión de sus supuestos y prácticas. Incorporar esta perspectiva al estudio de una infraestructura requiere establecer las relaciones históricas de apropiación, clasificación y subordinación pertinentes para el caso. Esa investigación histórica es necesaria para justificar la interpretación de un resultado técnico en términos de colonialidad.

En el espacio iberófono, la dependencia de proveedores externos puede coexistir con desigualdades internas de representación y reconocimiento. Una política que fortalezca capacidades en español o portugués debe examinar también las relaciones entre comunidades y las condiciones de su participación. Los microcasos analizados permiten localizar decisiones relevantes para ese examen. La observación de procesos comunitarios y el contraste de la aplicación conjunta de las tres condiciones quedan como tareas pendientes de investigación en infraestructuras concretas. Esa agenda incluye estudios sobre corpus multilingües, traducción automática entre variedades del español y del portugués, gobernanza comunitaria de datos lingüísticos y evaluación de modelos en lenguas criollas, indígenas o subrepresentadas. Su desarrollo permitirá contrastar la utilidad del marco para comprender la automatización lingüística como un problema de comunicación, poder y legitimidad epistemológica. Desarrollar tecnologías lingüísticas propias exige decidir, en el nivel de corpus, estándares y

mediaciones, qué formas de lengua, comunidad y conocimiento queremos que esas tecnologías hagan posibles y quién participa en esa decisión.

6. REFERENCIAS

Baack, S. (2024). A critical analysis of the largest source for generative AI training data: Common Crawl. En *Proceedings of the 2024 ACM Conference on Fairness*, *Accountability*, *and Transparency*, Association for Computing Machinery, 2199-2208. https://doi.org/10.1145/3630106.3659033

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

Bender, E. M., Gebru, T., McMillan-Major, A. y Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? En *Proceedings of the 2021 ACM Conference on Fairness*, *Accountability*, *and Transparency*, Association for Computing Machinery, 610-623. https://doi.org/10.1145/3442188.3445922

Birhane, A. y Guest, O. (2021). Towards decolonising computational sciences. *Kvinder*, *Køn & Forskning*, *29*(1), 60-73. https://doi.org/10.7146/kkf.v29i2.124899

Couldry, N. y Mejias, U. A. (2019a). Making data colonialism liveable: How might data's social order be regulated? *Internet Policy Review*, *8*(2), 1-16. https://doi.org/10.14763/2019.2.1411

Couldry, N. y Mejias, U. A. (2019b). *The costs of connection: How data is colonizing human life and appropriating it for capitalism.* Stanford University Press. https://doi.org/10.1515/9781503609754

Crawford, K. (2021). *Atlas of AI: Power*, *politics*, *and the planetary costs of artificial intelligence*. Yale University Press.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M. y Gardner, M. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. En M. -F. Moens, X. Huang, L. Specia y S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1286-1305). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.98

Fanon, F. (2009). *Piel negra*, *máscaras* blancas. Akal. (Obra original publicada en 1952).

Feenberg, A. (1999). *Questioning technology*. Routledge.

Floridi, L. (2019). What the near future of artificial intelligence could be. *Philosophy & Technology*, 32, 1-15. https://doi.org/10.1007/s13347-019-00345-y

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H. y Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92. https://doi.org/10.1145/3458723

Glissant, É. (1997). *Poetics of relation* (B. Wing, Trad.). University of Michigan Press.

Gordon, L. R. (2006). *Disciplinary decadence: Living thought in trying times*. Paradigm Publishers.

Lakew, S. M., Erofeeva, A. y Federico, M. (2018). Neural machine translation into language varieties. En O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi y K. Verspoor (Eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 156-164). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6316

Nee, J., Smith, G. M., Sheares, A. y Rustagi, I. (2022). Linguistic justice as a framework for designing, developing, and managing natural language processing tools. *Big Data & Society*, *9*(1), 1-6. https://doi.org/10.1177/20539517221090930

Papineni, K., Roukos, S., Ward, T. y Zhu, W. -J. (2002). BLEU: A method for automatic evaluation of machine translation. En P. Isabelle, E. Charniak y D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

UNESCO. (2022). *Recommendation on the ethics of artificial intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000380455

UNESCO. (11 de diciembre de 2023). *New report and guidelines for indigenous data sovereignty in artificial intelligence developments*. https://www.unesco.org/en/articles/new-report-and-guidelines-indigenous-data-sovereignty-artificial-intelligence-developments

Verbeek, P.-P. (2008). Cyborg intentionality: Rethinking the phenomenology of human-technology relations. *Phenomenology and the Cognitive Sciences*, 7, 387-395. https://doi.org/10.1007/s11097-008-9099-x

Visvanathan, S. (2009). The search for cognitive justice. *Seminar*, 597. https://www.india-seminar.com/2009/597/597_shiv_visvanathan.htm

Wynter, S. (2003). Unsettling the coloniality of being/power/truth/freedom: Towards the human, after Man, its overrepresentation—An argument. *CR: The New Centennial Review*, *3*(3), 257-337. https://doi.org/10.1353/ncr.2004.0015

Zampieri, M., Nakov, P. y Scherrer, Y. (2020). Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, *26*(6), 595-612. https://doi.org/10.1017/S1351324920000492

CONTRIBUCIONES DE AUTORES, FINANCIACIÓN Y AGRADECIMIENTOS

Financiación: Esta investigación no recibió financiamiento externo.

Agradecimientos: Este trabajo se inscribe en el marco de C3T. Agradezco a la CPA, a Óscar Guardiola-Rivera y a Stefan Baumgarten el diálogo y las reflexiones compartidas sobre automatización.

Conflicto de intereses: no existen

AUTOR:

Manuel De-la-Cruz-Recio
Universidad de Salamanca. Departamento de Traducción e Interpretación. España
Doctor por la Universidad de Granada y profesor del Departamento de Traducción e Interpretación de la Universidad de Salamanca. Su investigación articula el pensamiento complejo, la semiótica y la teoría crítica de la tecnología para estudiar la traducción, la interpretación y la automatización lingüística. Analiza las implicaciones epistemológicas, éticas y sociales de la inteligencia artificial y las transformaciones de la agencia humana en la mediación intercultural. Ha realizado estancias académicas en Heidelberg, Maguncia y Graz y publicado en editoriales como Routledge y John Benjamins. Participa en C3T y desarrolla iniciativas de innovación docente y transferencia del conocimiento, entre ellas LinguaCast y Don de Lenguas, vinculadas a la comunicación oral, el multilingüismo y la formación de intérpretes.
manueldelacruzrecio@usal.es
Índice H: 1
Orcid ID: https://orcid.org/0000-0002-0686-6652
Google Scholar: https://scholar.google.com/citations?user=jDnIYN0AAAAJ
ResearchGate: https://www.researchgate.net/profile/Manuel-De-La-Cruz-Recio
Scopus: https://www.scopus.com/authid/detail.url?authorId=24472949800
Academia.edu: https://usal.academia.edu/Manueldelacruz

Revista de Comunicación de la SEECI
ISSN 2695 - 5156

Artículos relacionados

Delso Vicente, A. T., Carvajal Camperos, M. y Corral De La Mata, D. Ángel. (2024). La evolución del procesamiento del lenguaje natural y su influencia en la inteligencia artificial: Una revisión y líneas de investigación futura. *European Public & Social Innovation Review*, 10, 1-23. https://doi.org/10.31637/epsir-2025-782

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

Mayor-Rocher, M., del Pozo Huerta, C., Melero, N., Martínez, G., Grandury, M. y Reviriego, P. (2025). Es igual pero no es lo mismo: ¿Distinguen los LLMs las variedades del español? *Procesamiento del Lenguaje Natural*, 75, 137-146. https://doi.org/10.26342/2025-75-10

Puukko, O. (2023). Repensando los derechos digitales a través de los problemas sistémicos de la comunicación. *Revista Latina de Comunicación Social*, 82, 1-19. https://doi.org/10.4185/rlcs-2024-2044

Peña Cáceres, O., Sánchez-Rogel, E., Barros-Naranjo, J., Silva-Marchan, H., Espinoza-Nima, R. y Correa-Calle, T. (2025). Linguistic and Visual Patterns of ChatGPT in Higher Education: An Analysis of its Use in Undergraduate Theses. *VISUAL REVIEW. International Visual Culture Review Revista Internacional De Cultura Visual*, *17*(3), 265-278. https://doi.org/10.62161/revvisual.v17.5403