TRASLACIONES
Revista Latinoamericana de Lectura y Escritura
CÁTEDRA UNESCO para la lectura y escritura
Vol. 10 (20) diciembre - junio 2024 - ISSN 2362- 6194

RETOS FRENTE A LAS TECNOLOGÍAS DIGITALES DEL LENGUAJE. UNA PERSPECTIVA GLOTOPOLÍTICA

CHALLENGES FACING DIGITAL LANGUAGE TECHNOLOGIES. A GLOTOPOLITICAL PERSPECTIVE

Yvette Bürki[1]

yvette.buerki@unibe.ch

Institut für Spanische Sprache und Literaturen

Universität Bern

Suiza

RESUMEN

En este artículo se abre un espacio de reflexión crítica sobre las consecuencias sociopolíticas y económicas que conlleva la intervención de las nuevas tecnologías digitales del lenguaje en las prácticas lingüísticas y discursivas de los hablantes. Así, desde una perspectiva que combina la glotopolítica y el giro poshumanista, se tratan algunos de los retos que traen estas nuevas tecnologías digitales en la creación y recreación de jerarquías sociolingüísticas y que acarrean inequidad social. Se muestra el poderoso entramado económico que sostiene tecnológicamente a lenguas súpercentrales como el inglés y, en menor medida al español; y, por otro lado, se señalan los efectos negativos para lenguas fuertemente minorizadas y descapitalizadas como las amerindias. Finalmente, se abordan las consecuencias glotopolíticas para el español, entendida como lengua pluricéntrica. Se emplean para ello ejemplos concretos procedentes de dos tecnologías del lenguaje diferentes: los traductores automáticos y las asistentes de voz. Este artículo se ha concebido ante todo como un espigueo sobre el desarrollo y el uso de tecnologías digitales del lenguaje y su relevancia glotopolítica en relación al español y las lenguas amerindias, que merecen estudiarse con mayor calado en investigaciones futuras.

Palabras clave: Tecnologías digitales del lenguaje - Glotopolítica - Giro posthumanista - Inglés - Español como lengua pluricéntrica - Lenguas amerindias

ABSTRACT

This article opens a space for critical reflection on the socio-political and economic consequences of the intervention of new digital language technologies in the linguistic and discursive practices of speakers. In light of this, from a perspective that combines glotopolitics and the post-humanist turn, it addresses some of the challenges posed by these new digital technologies in the creation and recreation of sociolinguistic hierarchies that bring about social inequity. The powerful economic framework that technologically sustains supercentral languages such as English and, to a lesser extent, Spanish, is exposed; and, on the other hand, the negative effects for strongly minoritized and decapitalized languages such as Amerindian languages are pointed out. Finally, it deals with the glotopolitical consequences for Spanish, understood as a pluricentric language. Specific examples are used for this purpose from two different language technologies: automatic translators and voice assistants. This article has been conceived primarily as a deep dive into the development and use of digital language technologies and their glotopolitical relevance in relation to Spanish and Amerindian languages, which deserve to be studied in greater depth in future research.

Keywords: Digital language technologies - Glotopolitics - Posthumanist turn - English, Spanish as a pluricentric language - Amerindian languages

Recepción: 02-09-2023 **Aceptación:** 28-11-2023

INTRODUCCIÓN

En este artículo avanzamos reflexiones sobre las consecuencias glotopolíticas de las tecnologías digitales del lenguaje, como por ejemplo la traducción, la corrección, la generación automática de textos y los asistentes de voz, centrándonos para ello en el español y las lenguas amerindias. La glotopolítica es especialmente útil para abordar estos desarrollos y usos actuales porque es una perspectiva de análisis que pone el foco en las intervenciones políticas y económicas en el espacio del lenguaje y las ideologías lingüísticas que las sustentan, con distintas consecuencias para los hablantes (Arnoux y del Valle, 2010; Arnoux, 2014). Utilizo el término *tecnologías digitales del lenguaje* para referirme a los instrumentos de comunicación ideados y desarrollados para comunicar en Internet y la web como parte integral de nuestra interacción en la vida cotidiana. Estas nuevas tecnologías digitales crean, por un lado, nuevas formas de homogenización y regulación lingüísticas como efecto de los modelos de lenguas dominantes, básicamente el inglés, que es la lengua que sirve tanto como fuente para la extracción y procesamiento de datos como para la programación de la ingeniería digital requerida (Arnoux y Lauria, 2023; Schneider, 2022); por otro lado, dichas tecnologías crean y recrean jerarquías sociolingüísticas como producto de ideologías lingüísticas dominantes. Al (re)producir estas jerarquías sociolingüísticas, también (re)producen formas de poder, las cuales pueden ir aparejadas con formas de dominación poscolonial y, en consecuencia, con manifestaciones de inequidad social.

El artículo está dividido en tres partes. En la primera, explicamos cómo funcionan estas tecnologías basadas en el procesamiento natural del lenguaje (PNL). En la segunda parte, nos centramos en las consecuencias glotopolíticas de las prácticas tecnológicas digitales comerciales con especial atención en el español y las lenguas amerindias. Expondremos de qué manera las formas de procesamiento de lenguaje natural no solo conllevan la reafirmación de jerarquizaciones sociolingüísticas existentes, recreando .sesgos ideológicos producto de dichas jerarquías, sino también la simplificación y la homogenización lingüísticas que, en último término, constituyen formas de regulación política del lenguaje. En la tercera parte, abordamos el influjo y las consecuencias del desarrollo de tecnologías digitales del lenguaje para las distintas variedades del español y de qué manera interactúan en este escenario diferentes agentes de poder político y económico de la lengua española. En especial, me ocupo aquí del Proyecto Lengua Española e Inteligencia Artificial (LEIA) que tiene la Real Academia Española (RAE) en cooperación con Telefónica y grandes corporaciones tecnológicas (Amazon, Google y Microsoft) con el objetivo de tematizar formas de poder en las que se entrelazan lo comercial con lo político. Por último, a manera de ejemplos de cómo estas corporaciones presentan la variación del español, analizamos el funcionamiento de herramientas tecnológicas comerciales en el ámbito escrito (traductores automáticos) y en el ámbito

oral (las asistentes de voz). Hemos optado por estas dos tecnologías porque ejemplifican su funcionamiento en dos medios distintos (escrito vs. oral) y porque son dos tecnologías bastante extendidas entre las usuarias y los usuarios, tanto en el dominio privado, como empresarial. Sobre todo, las asistentes de voz están aumentando exponencialmente sus dominios de uso[2].

Con este artículo, nos proponemos contribuir a la reflexión crítica y eminentemente glotopolítica sobre las consecuencias de las tecnologías digitales del lenguaje desde la perspectiva de lo que se conoce como el giro posthumanista (Ferrando & Brito Ledesma, 2022[2013]; Pennycook, 2016; Schneider, 2021): una apuesta onto y epistemológica en la que el elemento no humano, como por ejemplo en este caso específico, lo tecnológico, se entiende como una extensión de lo humano y no como una forma nítidamente separada de ello. En este contexto, lo humano no se plantea, entonces, como un agente autónomo, sino como parte de un amplio sistema de relaciones con el que comparte acciones y agencias. En un mundo altamente globalizado, en el que el orden de la economía política neoliberal es el imperante y las tecnologías digitales del lenguaje son omnipresentes, dicha forma de abordar las interrelaciones entre lenguaje y tecnología arrojan luz sobre las subjetividades actuales y de qué manera estas indicializan las jerarquías sociopolíticas en el terreno lingüístico.

1. Ontología de las nuevas tecnologías del lenguaje digital

Las tecnologías digitales del lenguaje son un producto del PNL, que, junto con los campos afines del reconocimiento automático del habla y la conversión de texto en habla, constituyen un subconjunto de conocimientos de lo que se entiende como inteligencia artificial (IA). Este campo de conocimiento tiene por objetivo dotar a las computadoras de la capacidad de entender y generar lenguaje humano de forma muy similar a como lo hacen los seres humanos (Bommasani et al., 2021). Entre los servicios que brindan estas tecnologías pueden mencionarse:

- corrección automática de textos (por ejemplo, el incorporado en el programa Word o Language Tool que trabaja con Google Docs o el programa Enclave de la RAE);
- traducción automática de textos (por ejemplo, Google Translate, Translator de Microsoft o DeepL);
- sistemas de transcripción automatizados (por ejemplo, Transcribe, Whisper AI o Sonix);
- generación automática de textos (por ejemplo, sistemas de mensajería automática como SMS y WhatsApp y correos electrónicos);

- herramientas conversacionales utilizadas para automatizar interacciones (por ejemplo, los programas que permiten interactuar a los usuarios mediante una interfaz de chat o los más sofisticados como Siri, Alexa o Cortana);
- análisis de sentimientos (por ejemplo, reconocimiento y categorización de palabras asociadas a emociones positivas o negativas, como ofrecen los softwares Idiomatic, MonkeyLearn o Lexalytics)[3];
- categorización automática de textos (por ejemplo, descripciones de productos para plataformas de mercado en línea o clasificación de grandes masas de documentos corporativos para ser localizados y compartidos, como permiten los softwares Cogito, Pangeanic o TextCortex);
- localización (por ejemplo, adaptación de textos, contenidos digitales o software a contextos culturales específicos);
- aprendizaje de lenguas extranjeras con programas automatizados (por ejemplo, Babbel, Duolingo o Rosetta Stone).

Como vemos a partir de estos ejemplos, en tanto que hablantes del español o del inglés, hacemos uso constante de tecnologías digitales del lenguaje, aunque esto no puede afirmarse para hablantes de muchas otras lenguas del mundo, básicamente del Sur Global (Bommasani et al., 2021; Bustamante et al., 2020; Ravindran, 2023; entre otros). Pero teniendo en cuenta estas importantes salvedades glotopolíticas, estos espacios y prácticas digitales son influyentes y formativos para una parte importante de la población mundial (Kelly-Holmes, 2019), por lo que ejercen *de facto* una acción normativa homogeneizadora sobre las prácticas lingüísticas que, a su vez, muestran la existencia de jerarquías lingüísticas en la arena sociopolítica global. Antes de entrar a tratar este tema en las siguientes secciones, ofreceremos a continuación una breve panorámica con el objetivo de contextualizar cronológicamente el desarrollo de estas tecnologías.

Los primeros intentos de hacer que las computadoras generaran lenguaje humano se basan en la programación manual en código informático para el procesamiento de elementos léxicos y reglas gramaticales del lenguaje humano (Moreno Sandoval, 2019). Estos sistemas, denominados *procesamiento simbólico del lenguaje natural* (PSLN) o *comprensión del lenguaje natural* (CLN), surgen a partir de las teorías chomskianas en los años 50, en el ámbito de la traducción automática.

Aunque dichos sistemas sientan las bases para aplicaciones de PNL más complejas, pronto se descubren sus limitaciones. Además de requerir mucho esfuerzo para crear las reglas, su mayor desventaja radica en su incapacidad para escalar, generalizar y adaptarse a la naturaleza compleja y dinámica del lenguaje humano (Fergus & Chalmers, 2022).

Estos problemas llevan a un cambio de paradigma, iniciado a principios de los años 90 del siglo pasado: la introducción de métodos estadísticos. El empleo de la estadística constituye un gran avance en la investigación de la PNL, ya que permite a las computadoras aprender de los datos y mejorar su rendimiento. Los métodos estadísticos requieren recoger grandes colecciones de ejemplos y datos para poder calcular, a partir de ellos, las frecuencias de diferentes unidades lingüísticas (letras, palabras, oraciones) y su probabilidad de aparecer en un contexto determinado. Calculando esta probabilidad, se puede predecir cuál será la siguiente unidad en un contexto dado, sin necesidad de recurrir a reglas gramaticales explícitas (Moreno Sandoval, 2019). Como resultado, los métodos estadísticos se han hecho cada vez más populares en la investigación de la PLN y se utilizan ampliamente en aplicaciones como la traducción automática, el análisis de sentimientos y la clasificación de textos (Cuantum Tecnologies, 2023). Recordemos que las tecnologías digitales del lenguaje no tienen por objetivo programar algoritmos para generar o descodificar reglas gramaticales del lenguaje humano, sino interpretar y generar lenguaje humano para "ofrecer soluciones parciales a problemas de la vida real" (Moreno Sandoval, 2019, p.81), como traducir textos, buscar información en Internet o enviar mensajes electrónicos. Precisamente la programación de algoritmos para que las máquinas pudieran generar lenguaje humano constituye el inicio del aprendizaje automático (ing. *machine learning*) cuyas características residen en su capacidad para aprender de los datos y mejorar su rendimiento a lo largo del tiempo sin necesidad de una programación explícita mediante reglas realizadas por las personas.

Ahora bien, aunque los métodos estadísticos pueden captar patrones más complejos que los basados en reglas gramaticales, se necesita aún mucho esfuerzo de reajuste manual, sobre todo porque no manejan bien los aspectos más complejos del lenguaje, como por ejemplo, las expresiones idiomáticas, la ambigüedad y el contexto (Cuantum Tecnologies, 2023). Debido a ello, a partir de 2010 en adelante, cada vez más estos cálculos estadísticos de correlaciones de palabras se realizan mediante técnicas de redes neuronales artificiales (algoritmos interrelacionados), lo que significa que se utilizan múltiples algoritmos interconectados —las denominadas redes neuronales— para analizar estadísticamente enormes conjuntos de datos lingüísticos (los denominados *big data*), con escasa intervención humana (Fergus & Chalmers, 2022). Para ello, se necesita menos entrenamiento en la anotación y el tratamiento de datos en una fase de pre-entrenamiento, ya que, gracias al aprendizaje automático profundo, la misma arquitectura de red neuronal puede utilizarse para muchas aplicaciones (Bommasani et al., 2021). En contraposición, el costo computacional es elevadísimo debido a los complejos requisitos de ingeniería requeridos. En general, los enfoques que hacen uso de redes neuronales también se denominan tecnologías de aprendizaje profundo

(*deep learning*) y se asocian con la "inteligencia artificial" en el discurso no especializado (Schneider, 2022, pp.368-369).

En realidad, estos nuevos sistemas neuronales actúan mediante principios de agrupación guiados por la enorme masa de datos que los alimentan (Moreno Sandoval, 2019), lo cual les permite generar textos mediante procesos de generalización. Esto significa que los programadores humanos no deciden necesariamente qué características se consideran relevantes y que los conjuntos de datos no tienen que ser etiquetados por humanos (por ejemplo, según las partes del discurso) antes de que los algoritmos informáticos empiecen a calcular las relaciones entre palabras. Un modelo de algoritmos que funciona de esta manera es el de soporte vectorial, que se basa en el aprendizaje y generación automática de secuencias de signos de acuerdo con la probabilidad de su ocurrencia en determinado contexto. Programas como *ELMo* (2018), *Representations from Transformers* (2019), *BERT* (2019), *MarIA* (2022)[4] para el español y las series GPT actúan sobre la base de estos modelos (Cuantum Tecnologies, 2023; Gutiérrez-Fandiño et al., 2021; Sennrich, 2023). Con respecto a otros modelos anteriores, lo que los hace enormemente potentes es que pueden producir textos coherentes a gran escala. Además, en su última generación (por ejemplo, ChatGPT4) han sido entrenados mediante diálogos, lo cual aumenta su capacidad de interacción con las personas, de quienes, a su vez, aprenden (Sennrich, 2023).

Si observamos los siguientes datos, vemos cómo han ido incrementándose no solo la cantidad de datos para el entrenamiento en PNL, sino la multiplicación de redes neuronales, lo que va aparejado con los enormes costos de su procesamiento.

Tabla 1

Modelos vectoriales de acuerdo a sus dimensiones, datos y costos de entrenamiento

Modelo	Dimensiones de la red (parámetros empleados)	Datos de entrenamiento	Costos de entrenamiento
ELMo (2018)	93 millones	800 millones	900 USD
BERT (2018)	340 millones	3'300 millones	7000 USD
GPT-2 (2019)	1.500 millones	10.000 millones	50.000 USD
GPT-3 (2020)	175.000 millones	500.000 millones	4.600.000 USD
GPT-4 (2023)	1.8 trillones	300 billones	63.000.000 USD

Nota: tabla propia elaborada de acuerdo con Sennrich (2023) y The decoder (2023)

Ahora bien, como explica Sennrich (2023), desde el punto de vista técnico, estos programas, al estar basados en principios de categorización y predicción, son también "sorprendentemente estúpidos" (mi traducción), puesto que no distinguen entre hechos verdaderos y ficticios, entre hechos pasados y actuales y, sobre todo, —y de allí también su peligro—, detrás de los textos que generan no hay sentimientos, ni intenciones, ni valores morales, puesto que lo que los mantienen son modelos probabilísticos que generan palabras en función a los contextos (Chomsky et al., 2023). Pero sí hay intenciones sociopolíticas y económicas detrás de las entidades humanas que producen estas herramientas tecnológicas y que están impulsando implícitamente, y a través de las prácticas, nuevos procesos normativos de la comunicación y de las lenguas.

2. Prácticas de las tecnologías digitales y consecuencias glotopolíticas

La glotopolítica se define como el estudio de los instrumentos lingüísticos entendidos como intervenciones en el espacio público del lenguaje asociadas con ideologías lingüísticas en estrecha vinculación con las condiciones sociohistóricas en las que estos se producen y circulan (Arnoux y Lauria, 2023). Desde esta perspectiva de estudio, se ha abordado sobre todo la dimensión metadiscursiva de los instrumentos lingüísticos, como por ejemplo aparece en las gramáticas, los diccionarios, los tratados ortográficos o los manuales de texto escolar (Arnoux, 2014 y 2016). En el caso de las tecnologías digitales, la intervención reguladora se da más bien a partir de las propias prácticas, que son las que de manera implícita muestran los modelos de lengua más "aceptados", reflejando de esta forma los lugares y las ideologías lingüísticas de poder (del Valle, 2017). A continuación, ahondamos en determinados factores que influyen en estas intervenciones sociopolíticas y económicas en el plano del lenguaje y su repercusión en el español y en las lenguas amerindias.

2.1 El poder del inglés

En primer lugar, es insoslayable el papel que desempeña el inglés en la programación y difusión de las tecnologías digitales del lenguaje. Esto tiene que ver con la propia historia de la ciencia. Como señalan Danet y Herring (2003, citado en Kelly-Holmes, 2019), los primeros planificadores de Internet eran en general estadounidenses, e implícitamente solo tenían en mente una lengua: el inglés; no previeron, por lo tanto, los problemas que podrían surgir cuando hablantes de otras lenguas buscaran también comunicarse en línea. Así pues, la ideología y la cultura monolingüe (en inglés) se convirtieron en parte de la cultura de Internet y, más tarde, de la web. Este es un punto importante, porque como indica Kelly-Holmes (2019), "los avances tecnológicos se producen en espacios

ideológicos y culturales concretos, y la forma de esos avances tecnológicos lleva la impronta de esas normas culturales e ideológicas" (p.27, traducción propia).

De acuerdo con *Internet World Statistics*, en el año 2020 el inglés fue la lengua más empleada en la web, con 1.186.451.052 personas, lo que corresponde al 25.9% de la población mundial de usuarios/as. En segundo lugar, se encuentra el chino mandarín, con 888.453.068 personas que lo emplean, correspondiente al 19.4% de la población mundial de usuarios/as. En tercer lugar, pero bastante más abajo en niveles de porcentajes de uso, figura el español, con 363.684.593 personas, lo cual se traduce en un 7.9% del número total de usuarios/as. Hay que tener en cuenta además que muchos de estos usuarios/as no tienen el inglés, el chino o el español como L1, sino que acuden a estas lenguas por su disponibilidad en la web. Este hecho significa que existe la posibilidad de entrenamiento en PNL para estas lenguas. El aspecto técnico no es baladí, ya que la gran mayoría de los lenguajes de programación no solo están escritos en inglés, sino que se basan en la lengua inglesa (Kelly-Holmes, 2019). En 2019, por ejemplo, la mayoría de las herramientas de PNL se crearon para el inglés (Ravindran, 2023). Dado que las grandes compañías tecnológicas como Google, META, Apple, Amazon y grandes consorcios telefónicos como A&TT y Huawei utilizan efectivamente el inglés como lengua hegemónica de la tecnología y son además las que cuentan con los recursos económicos que se necesitan para llevar a cabo los procesos de entrenamiento y transformación de datos —como muestra la Tabla 1—, el poder del inglés como lengua virtual se perpetúa.

Por otro lado, el hecho de que se entrenen mediante aprendizaje automático las máquinas con corpus de base en inglés tiene repercusión en las estructuras de las lenguas mismas. Debido a la propia tecnología del procesamiento automático de datos que, como ya he expuesto, tienen como objetivo generar textos a gran escala mediante métodos probabilísticos que detectan generalizaciones (cfr. sección 1, *supra*), se desarrollan automáticamente tendencias hacia la homogenización y simplificación de las estructuras lingüísticas (Schneider, 2022). Vanmassenhove et al. (2019; 2021) demuestran los efectos homogeneizadores y simplificadores de los algoritmos en la traducción automática del inglés al español en la pérdida de la riqueza léxica, un aspecto que va más allá de las cuestiones relacionadas con la morfología de género, ya tratadas en artículos anteriores (Vanmassenhove et al., 2021; Passban et al., 2018). Así, por ejemplo, si el material en español original muestra que la palabra *picture* puede traducirse por una variedad bastante apreciable de palabras como 'imagen', 'imágenes', 'visión, 'foto', fotografía', fotografías, 'fotos', la traducción automática al inglés da como resultado más frecuente solo 'imagen', incluso con mayor frecuencia que en los datos originales, al tiempo que las palabras menos frecuentes tienden a perderse (Vanmassenhove et al.,

2019 p.229). Lo mismo se observa con la palabra *happen*, que en el material original se presenta con diferentes variantes léxicas como 'ocurrir', 'suceder', 'pasarse', 'acontecer', 'pasarse' pero que en la traducción automática gana 'ocurrir' en frecuencia relativa a costa de las opciones menos frecuentes (Vanmassenhove et al., 2019, p.230). En Vanmassenhove et al. (2021) se profundizan estos experimentos tanto en el plano léxico como morfológico, ya no solo para el español, sino también para el francés —ambas lenguas morfológicamente más ricas que el inglés. Los resultados a los que se llegan van en consonancia con los obtenidos en 2019, mostrando que los datos de entrenamiento originales tienen más diversidad léxica y morfológica en comparación con las producidas por la traducción automática.

No solo las estructuras y el léxico de las lenguas se ven permeadas por la traducción automática al inglés. En otros ámbitos, programas de PNL de última generación como ChatGPT 3 y 4 producen tipos de textos de acuerdo a patrones de la cultura textual anglosajona. Si, por ejemplo, se le encarga a ChatGPT producir un artículo académico lo hará reproduciendo patrones discursivos y estilísticos difundidos y naturalizados como modélicos en revistas indexadas, que, a su vez, toman como referencia los parámetros discursivos del inglés "como lengua hegemónica de la ciencia" (Arnoux y Lauria, 2023, p.138; Bein, 2020; Lara, 2015 y Navarro et al., 2022). En este sentido, el inglés ejerce un impacto regulador en las propias tradiciones discursivas del español que, en último término, son parte de su tradición cultural. Estas homogeneizaciones como producto de procesos de generalización de algoritmos basados en la preminencia del inglés en los datos como en las tecnologías de procesamiento, constituyen sesgos sociolingüísticos (cfr. Blodgett et al., 2020) en la medida en que reafirman y recrean jerarquías lingüísticas en las que los patrones del inglés constituyen el modelo de base en la difusión de normas discursivas globalizadas, lo cual pone en desventaja sociopolítica y económica a otras lenguas.

2.2 Otras lenguas de poder

Ahora bien, como muestra *Internet World Statistics* de 2020 (cfr. sección 1, *supra*), otras lenguas también han ido avanzando en la elaboración de tecnologías digitales del lenguaje. De acuerdo con esta fuente, las 10 lenguas más usadas en la Red para marzo de 2020 son las siguientes:

Tabla 2

Las diez lenguas más usadas en la web (marzo 2020) según Internet World Statistics (https://www.internetworldstats.com/stats7.htm)

Top Ten Languages in the Internet	World Population for this Language (2021 Estimate)	Internet Users by Language	Internet Penetration (% Population)	Internet Users Growth (2000 - 2021)	Internet Users % of World (Participation)
English	1,531,179,460	1,186,451,052	77.5 %	742.9 %	25.9 %
Chinese	1,477,137,209	888,453,068	60.1 %	2,650.4 %	19.4 %
Spanish	516,655,099	363,684,593	70.4 %	1,511.0 %	7.9 %
Arabic	447,572,891	237,418,349	53.0 %	9,348.0 %	5.2 %
Portuguese	290,939,425	171,750,818	59.0 %	2,167.0 %	3.7 %
Indonesian / Malaysian	306,327,093	198,029,815	64.6 %	3,356.0 %	4.3 %
French	431,503,032	151,733,611	35.2 %	1,164.6 %	3.3 %
Japanese	126,476,461	118,626,672	93.8 %	152.0 %	2.6 %
Russian	145,934,462	116,353,942	79.7 %	3,653.4 %	2.5 %
German	98,654,451	92,525,427	93.8 %	236.2 %	2.0 %
Top 10 Languages	5,273,725,132	3,525,027,347	66.8 %	1,188.2 %	

No obstante, este crecimiento es parcial, ya que como vemos en la tabla, han crecido sobre todo las grandes lenguas que, según De Swaan (2013), se definen como súpercentrales debido a que funcionan como lenguas centrales o lenguas francas o como lenguas con gran número de hablantes y/o gran poder económico (Kelly-Holmes, 2019). En una entrada del blog del Instituto de Ingeniería del Conocimiento[5] dedicado al PNL, y firmado por el lingüista computacional Antonio Moreno, se lee lo siguiente: "Virtualmente, cualquier lengua humana puede ser tratada por los ordenadores. Lógicamente, limitaciones de interés económico o práctico hace que solo las lenguas más habladas o utilizadas en el mundo digital tengan aplicaciones de uso". Interesante es el empleo del modalizador epistémico *lógicamente*, que hace del contenido locutivo del enunciado un hecho naturalizado y que nos muestra que, en efecto, la presencia y difusión de las lenguas en el campo digital conlleva poder, tanto sociopolítico como económico. De allí también se explica el fuerte crecimiento que han experimentado lenguas entre 2000-2021 como el chino (2,650.4%) y el ruso (3,653.4%), pertenecientes a países con enormes ambiciones imperialistas, pero también, lenguas como el árabe (9,348.0%), el portugués, el español

y el francés, lenguas excoloniales que cuentan con gran número de hablantes no solo en Europa, sino en las Américas y en África.

2.3 La ideología capitalista y la brecha entre el Norte y el Sur Global

Efectivamente, la lógica del capitalismo tardío que caracteriza al mundo globalizado actual es la que proporciona una estructura de incentivos en las decisiones que se toman con respecto a la elección de las lenguas en las que se desarrollan instrumentos tecnológicos del lenguaje. Siguiendo esta lógica, la enorme inversión financiera que supone el desarrollo de programas de PNL, tiene que poder rentabilizarse económicamente a corto o mediano plazo, lo cual no es posible en sociedades con pocos recursos y escasas posibilidades de crecimiento económico. Al igual que la industria farmacéutica, tiene pocos incentivos para dedicar recursos significativos a la investigación y el desarrollo de tratamientos contra la malaria, porque la gente con muy pocos medios económicos no puede permitirse los medicamentos, la industria tecnológica tiene pocos incentivos para proporcionar las fuentes financieras necesarias destinadas al diseño de tecnologías para elevar cualitativamente las condiciones de vida de sociedades más pobres y periféricas (Reich et al., 2021). Esta comparación también pone de manifiesto la importancia del mercado lingüístico en términos de Bourdieu (1985), donde los productos lingüísticos se capitalizan y adquieren sus valores de manera contextualizada y relacional. Esto explica en parte por qué la inmensa mayoría de las casi 7.000 lenguas del mundo carece de datos, herramientas o técnicas de PNL, lo que las convierte en lenguas “con pocos recursos tecnológicos”, en contraste con un puñado de lenguas “con muchos recursos”, como el inglés, el chino mandarín, el español, el francés o el alemán (Ravindran, 2023, p.262).

Tres son los problemas a los que se enfrentan estas lenguas minorizadas, la mayoría de estas habladas en el Sur Global: en primer lugar, no hay suficientes datos a disposición para que los modelos de IA se entrenen. Aparte de la cuestión socioeconómica que adelanté arriba (cfr. sección 1, supra), esto también se debe al hecho de que durante siglos estas lenguas han sido básicamente orales, y han estado subordinadas para contextos institucionales, como por ejemplo los educativos, a las grandes lenguas, la mayoría de estas excoloniales. Este no es un problema puramente técnico, sino ante todo sociolingüístico y glotopolítico ya que bajo los regímenes coloniales se disuadía a los pueblos indígenas de utilizar sus lenguas, sobre todo en la escritura (Ravindran, 2023). En segundo lugar, los modelos algorítmicos están entrenados básicamente con corpus en inglés. Por ejemplo, las herramientas que „tokenizan“ o separan las frases en inglés en palabras individuales no funcionan bien para las lenguas aglutinantes (Ravindran, 2023), como sucede con muchas lenguas amerindias (Bustamante et al., 2020). Tampoco los modelos entrenados con corpus multilingües parecen muy útiles para representar

aspectos de lenguas que son drásticamente distintas del inglés o para las que se dispone de pocos recursos lingüísticos (Bommasani et al., 2021).

Por último, está la cuestión del acceso a esta ingeniería electrónica, ya que la mayoría de los modelos más avanzados de ingeniería computacional para el PNL se ha desarrollado en contextos comerciales sin acceso abierto y, dada la enorme inversión de recursos financieros que requiere la creación, entrenamiento e implementación de estos modelos, las universidades u organismos sin fines de lucro no tienen las capacidades para marchar al mismo ritmo que las grandes tecnológicas comerciales (Bommasani et al., 2021). De esta manera, se reproducen dependencias, jerarquías lingüísticas y formas veladas de colonización lingüística.

Hay que mencionar en este punto que una cosa es el aprovechamiento de redes sociales y de aplicaciones para la revitalización lingüística que, efectivamente está sufriendo un verdadero *boom* con respecto a las lenguas indígenas americanas (Alvarado, 2022 y Coronel Molina, 2019, por ejemplo), y otra cosa es desarrollar tecnología computarizada profunda para alcanzar la representación de estas lenguas en la web (Zariquiey, 2023) que, a fin de cuentas, es lo que proporcionaría independencia sociolingüística en el terreno digital mediante la incorporación, en iguales condiciones frente a las lenguas dominantes, del tipo de tecnología que a hablantes de lenguas minorizadas les gustaría encontrar en sus lenguas: generación de chats, reconocedores de voz o corrección automática de textos (Zariquiey, 2023). Ilustraré estas cuestiones a partir del PNL de lenguas amazónicas del Perú. En este país, además del español, que es la lengua más empleada, se hablan 48 lenguas indígenas agrupadas en 19 familias, casi todas situadas en la región amazónica del país (Bustamante et al., 2020). Estas lenguas han sido de empleo básicamente oral y solo en el marco de su reciente reconocimiento oficial en los últimos 10 años (Sullón Acosta, 2013) existen documentos —aunque aún pocos— escritos en lenguas indígenas; se trata sobre todo de los que vienen desarrollándose en el marco de los programas de educación bilingüe. Por otro lado, desde hace algunos años, se vienen realizando esfuerzos para el desarrollo de herramientas computacionales en lenguas amazónicas como el shipibo conibo, el ashaninka, el yanesha, el yank y el yine, lenguas que según GlottoScope (Hammarström et al., 2021), se encuentran en peligro (por ejemplo, Alva y Oncevay, 2017; Bustamante et al., 2020; Ortega et al., 2020; Oncevay et al., 2022). Lo que muestra el trabajo en PNL en estas lenguas es que, en efecto, la poca existencia de datos no siempre puede paliarse empleando modelos automatizados para generalización y empleo escalar de datos, como ocurre con los modelos de segmentación y parcelación ya existentes en inglés, debido a la riqueza morfológica de estas lenguas (fuertemente aglutinantes y sintéticas). Por el contrario, esta complejidad lingüística requiere, por ejemplo, la utilización de filtros mediante procesos manuales (Bustamante et al., 2020) o la creación de modelos de evaluación cuyos costos son elevados (Oncevay et al., 2022).

La poca disponibilidad de estas tecnologías digitales en lenguas indígenas minorizadas implica, entonces, que sus hablantes acudan a otras lenguas —como por ejemplo el español—, dada la vinculación actual de este tipo de herramientas a la economía global, lo cual (re)crea y perpetúa las dependencias sociolingüísticas y sociopolíticas.

Si estos son los problemas con los que nos encontramos con lenguas minorizadas que refuerzan, en el terreno virtual, la dependencia de estas lenguas —y sobre todo de sus hablantes— de las grandes lenguas, o lo que De Swaan (2013) llamaba lenguas supercentrales, entre las variedades del español también encontramos una serie de indicadores ligados al desarrollo e implementación de herramientas del lenguaje digital que privilegian determinadas variedades del español frente a otras, como expondremos en el siguiente apartado, y que también conllevan en el paisaje lingüístico virtual a la homogeneización del español, que como ya harto sabemos, es una lengua pluricéntrica.

3. Política lingüística e IA en español

En el último Congreso Internacional de la Lengua Española (CILE) de 2023, realizado en la Universidad de Cádiz, el panel "Lengua, inteligencia artificial e (in)dependencia tecnológica" trató desde diferentes perspectivas —tecnológicas, sociolingüísticas y sociopolíticas— el futuro del español en relación con la IA. Nunca mejor puesto el título, ya que estas "(in)dependencias" a la que refiere el título no solo se dan con respecto al inglés, sino a la variación misma del español, lo que incluye también las normas hispanas como conformantes de un español entendido como sistema pluricéntrico. En este marco, Virginia Bertolotti (2023) mencionó en su intervención dos cuestiones sociolingüísticas fundamentales por las cuales la representatividad de datos de las variedades del español destinados al tratamiento en IA debe ser tanto cuantitativa como cualitativamente equilibrada: la primera tiene que ver con razones de filiación e identidad, mientras que las segundas, con razones de ciudadanía. Las razones identitarias se explican por la propia naturaleza de la comunicación humana, pues si tenemos que interactuar a menudo con "alguien" es indudable que se nos hará más fácil y más agradable si asociamos a ese "alguien" como parte de nuestra comunidad. Las razones de ciudadanía tienen que ver con la accesibilidad de información que tienen los ciudadanos con respecto a sus derechos, lo que incluye los servicios y prestaciones públicas, como por ejemplo el servicio de información administrativa, pero también el sanitario, judicial, educativo, de protección laboral, etc. Dados los tiempos que corren, las informaciones sobre estos servicios se encuentran cada vez más en portales electrónicos públicos y, para hacerlas accesibles a la ciudadanía, deben ser transmitidas de manera lingüísticamente comprensible para esos ciudadanos; en otras palabras, en la variedad que les es familiar o conocida (Bertolotti, 2023).

Ahora bien, el español, dados su enorme difusión espacial y su complejo entramado social y cultural, tiene, como afirma Bertolotti (2023), "unos ejes de sofisticación variacional realmente considerables". Por otro lado, aunque los modelos de PNL actuales son bastante versátiles con los conocimientos que obtienen tras el entrenamiento con los datos, precisamente no está claro hasta qué punto pueden gestionar la variación lingüística en sus diferentes niveles (espacial, social y pragmático) (Bommasani et al., 2021). Aquí es entonces que los factores sociales y políticos influyen en la forma en que se considera y valora la variación lingüística, así como en el grado de representación de las distintas variedades en la investigación de la PNL.

Teniendo en cuenta este contexto, resulta glotopolíticamente interesante el discurso introductorio que sostuvo el director actual de la RAE, Santiago Muñoz Machado, en el encuentro sobre Lengua Española e Inteligencia Artificial (LEIA) en el año 2021 y en el que participaron representantes de las grandes tecnológicas (Alexa España, Google España, Microsoft y el director de la unidad global de consumo digital de Telefónica y director técnico del proyecto LEIA, Chema Alonso:

> Pretendemos que la inteligencia artificial hable español. Y que además lo hable bien. Que las máquinas parlantes hablen bien. Que traduzcan mejor. Que la manera de manejar nuestra lengua se adecúe a los cánones que ha establecido la RAE desde hace ya trecientos años. Los humanos de todo el mundo, cuando son hispanohablantes, se atienen a los criterios de la RAE en cuanto al léxico admisible, la gramática útil, la ortografía correcta. Lo hacen porque creen que la autoridad de la Real Academia Española (RAE) y la calidad de sus obras, no porque la Academia tenga un poder especial para imponerlo. Nos resulta muy difícil imponer esa autoridad a las máquinas que no nos entienden bien. Pero nos entienden mejor los dueños de las máquinas: las grandes corporaciones mundiales que las fabrican, y queremos convencerles de que realmente usen las herramientas de la RAE para enseñar, entrenar a las máquinas y hacerlas hablar conforme a ese canon que la RAE viene estableciendo y al que se atienen 600 millones de personas en el mundo. Los 700.000 millones de máquinas que producen y que hacen operaciones con la lengua tendrían igualmente que seguir esos criterios homologados, esos criterios que la RAE estableció. (Santiago Muñoz Machado, 2021)

Conforme leemos lo que señala el director de la RAE, surgen varias preguntas con relación a la lengua española. Una pregunta fundamental es qué significa "hablar bien" y en relación a qué variedad. En este sentido, es necesario recordar dos cuestiones importantes. En primer lugar, el hecho de que las variedades no son planas, sino que están compuestas por dimensiones espaciales, sociales y pragmáticas. Y, en segundo lugar, el hecho de que las y los hablantes de la lengua española comunican mediante "españoles nacionales" compuestos por "una jerarquía de normas reales, no necesariamente prescriptivas" (Lara, 2007, p.179) que contienen usos prestigiosos no siempre coincidentes entre ellos y que suelen apuntar a la existencia de estándares implícitos (Haas, 1982). Además, vale

recordar que las variedades, en términos de identidad, son siempre relacionales (Bucholtz y Hall 2005) y, en este sentido, son precisamente los aspectos sociolingüísticamente salientes los que pueden indicializar valor identitario entre los hablantes.

3.1 La variación escrita: los traductores automáticos

En esta línea de argumentación, la cuestión de lo que es el 'léxico admisible' y la 'gramática útil' en una lengua como la española, con un espectro variacional muy rico, pero por ello complejo, interpelan. Es cierto que los traductores automáticos han avanzado muchísimo. Por ejemplo, si se quiere traducir una frase como "I went with my bathrobe at school" al español, *DeepL* te da como traducción: 'Fui en bata a la escuela', con opción de emplear también las variantes 'albornoz' o 'batín' -sobre el par *albornoz/ bata*, (Caravedo, 2014). *Google Translate*, en cambio, solo te da 'bata' como única opción léxica, lo mismo que *Microsoft Translator*. Interesante resulta el hecho de que *bata,* que es una variante americana, sea la primera opción tanto en *Google Translate* como en *Microsoft Translator*, al tiempo que sorprende que sea solo *DeepL* que dé como opción la variante europea *albornoz.* Animada por este primer resultado, tanteé otras traducciones para léxico característico de determinadas variedades espaciales (y nacionales) del español utilizando tecnologías digitales del lenguaje, asegurándome, para ello, de desambiguar lo más posible los contextos para disminuir el margen de error. A continuación, la tabla resume los resultados:

Tabla 3

Traducciones y variantes ofrecidas por diferentes herramientas de traducción automática

Inglés	Esp. (DeepL)	Esp. (Google Translate)	Esp. (Microsoft Translator)
I went to **the swimming pool**	Fui a la **piscina**	Fui a la piscina	Fui a la **piscina**
Pass me the **butter** please	Pásame la **mantequilla**, por favor	Pásame la **mantequilla**, por favor	Pásame la **mantequilla**, por favor
Please, pass me the **pullover**	Por favor, pásame el **suéter** - opción: **jersey**	Por favor, pásame el **jersey**	Por favor, pásame el **jersey**
The bags are in the **trunk**	Las bolsas están en el **maletero** - opción: **baúl**	Las bolsas están en el **maletero**	Las bolsas están en el **maletero**

I need new **glases**	Necesito **gafas** nuevas - opción: **lentes** **monturas** **anteojos**	Necesito **gafas** nuevas	Necesito **gafas** nuevas
I bought a new **car**	He comprador un **coche** nuevo - opción: **automóvil** **vehículo** **auto**	Yo compré un **coche** nuevo	Compré un **auto** nuevo
I need new **jeans**	Necesito unos **vaqueros** nuevos - opción: **jeans**	Necesito **jeans** nuevos opción: **pantalones**	Necesito **jeans** nuevos
I have to buy **matches**	Tengo que comprar **cerillas** - opción: **fósforos**	Tengo que comprar **fósforos**	Tengo que comprar **fósforos**
My **computer** is broken	Mi **ordenador** está roto	Mi **computadora** está rota	Mi **computadora** está rota

El pequeño experimento arroja resultados interesantes en torno a lo que "entienden" las máquinas como variación. Así, en ninguno de los casos se da como opción para la traducción de *butter* 'manteca', que sería la preferida en la Argentina, por ejemplo, ni para la traducción de *swimming pool* 'pileta', que otra vez sería la opción preferida en la Argentina[6]. No olvidemos que en estos dos casos no solo hablamos de variantes dialectales, sino de palabras que pertenecen a la norma nacional de ese país. Tampoco se dan otras opciones, como por ejemplo 'alberca', que sería una variante para México. Al revés sucede con *pullover*, donde DeepL da como traducción 'suéter' y como segunda opción la variante europea 'jersey', mientras que los traductores de Google y Microsoft optan por la variante europea. En el caso de *car*, en cambio, se alinean DeepL y Google con 'coche', mientras que Microsoft da como única opción 'auto'. Otra vez, el único traductor automático que ofrece alternativas de traducción es DeepL. Interesante es la traducción de *trunk*, puesto que tanto DeepL, Google Translate como Microsoft Translator dan como traducción 'maletero' que corresponde a la variante europea; solo DeepL ofrece una variante más: 'baúl', vocablo que con ese significado se emplea en Argentina, Colombia, Cuba, Guatemala, Honduras y República Dominicana, según informa el Diccionario de la Lengua Española (DLE).

Algo semejante ocurre con *gafas*, que es la variante europea, siendo otra vez DeepL la única que da como alternativas *lentes* y *anteojos*, que son las empleadas en las variedades americanas, aunque el DLE sorprendentemente no las consigne con marcas diatópicas. Casos contrarios son las traducciones de *computer*, *jeans* y *matches,* donde Google Translate y Microsoft Translator prefieren la variante americana, 'computadora', 'jeans' y 'fósforos' respectivamente, que es la única que ofrecen como alternativa de traducción, mientras que DeepL, aunque consigna 'computadora' 'jeans' y 'fósforos' da como primera variante las formas europeas 'ordenador', 'vaqueros' y 'cerillas'. Este pequeño ejercicio muestra que las máquinas parecen partir de corpus diferentes y además bastante simplificados de la realidad lingüística hispana, corroborando las conclusiones a las que llegaban Vanmassenhove et al. (2019 y 2021) y Passban et al., (2018) en relación a la fuerte simplificación que sufre la variación mediante los programas de PNL (cfr. sección 2.1, *supra*). Solo DeepL registra de manera relativamente sistemática variantes y no siempre el criterio es el mismo. En esta línea, tampoco el criterio para la elección de una u otra variante parece quedar claro en los traductores automáticos de Google y de Microsoft puesto que no siempre es la opción europea o la americana la que prima. En resumen, y como ya apuntaban Vanmassenhove et al. (2019), las máquinas —sobre todo los traductores de Google y Microsoft— parecen estar "lost in translation" (p.230). Probablemente, esto se deba a que las masas de datos con las que se han alimentado los programas no han pasado por procesos de curaduría minuciosa, que son costosos tanto en cuanto al tiempo como al trabajo manual que se requiere. Así y todo, la repercusión que tienen en los usuarios/as a través de Internet, también en el ámbito de la educación (incluida ELE), que emplean estas herramientas a diario para traducir —muchas veces sin cuestionamiento o reflexión— no es desestimable y de allí que los repertorios que estas máquinas usan y su impacto en la (re)creación y/o de subjetividades y normatividades constituya un aspecto importante de análisis glotopolítico (Heyd y Schneider, 2019).

Si a través del proyecto LEIA, los traductores de Google y Microsoft encontrarán un camino más recto y "homologado" con los criterios policéntricos[7] de la RAE está por verse. Y, en todo caso, también está por verse cómo se interpretan y se ponen en práctica estos criterios, toda vez que no todos los corpus de la RAE están equilibrados en cuanto al espectro variacional[8] debido a la propia historia glotopolítica de la institución. Esto no quiere decir que no pueda paliarse el "defecto" y equilibrarse los corpus —de hecho, el CORPES es una buena muestra de ello— y que las buenas intenciones estén presentes. Asimismo, a pesar de los avances que se han hecho con el DLE en materia de variación, cabe preguntarse cómo influirían estos corpus en las opciones de traducción automática que ofrecen las grandes tecnológicas como Google y Microsoft, que participan en el proyecto LEIA, y más aún, de qué manera utilizará Telefónica estos *big data* que recibe de la RAE[9]. De acuerdo a la propia Telefónica, dos de sus decisiones operativas clave están en el desarrollo de tecnologías digitales, como reproducimos a continuación:

02. Lanzamiento de telefónica Tech:

> Gracias a esta nueva empresa se aspira a capturar el crecimiento del mercado de servicios digitales, con el objetivo de completar la oferta de conectividad a clientes corporativos. El foco inicial se ha centrado en una oferta de servicios sobre tres negocios: Ciberseguridad, Cloud e IoT/Big Data.

05. La aceleración de iniciativas de digitalización operativas en todas las unidades operativas del Grupo permitirá capturar sinergias de la simplificación y la automatización de los procesos. Adicionalmente, se ha abordado la reestructuración del Centro Corporativo, reduciendo su tamaño, para enfocarse en aquellas actividades que puedan aportar un valor diferencial a las operadoras, maximizar las sinergias y cristalizar el valor de la escala de Telefónica. (Telefonica.com, s. f.-a)

Telefónica española es líder en España y Brasil, es la tercera compañía de telefonía más grande en Alemania y, bajo la marca Virgin Media O2, es una de las más grandes en Gran Bretaña (telefónica.com). Como Grupo Telefónica, está presente en: Argentina, Chile, Ecuador, Perú, México, Uruguay, Venezuela y Centroamérica (telefonica.com), o sea en una gran parte de la Latinoamérica hispanohablante. Por su parte, Cristina Gallach Figueras (2023), actual presidenta del Proyecto estratégico para la Recuperación y Transformación Económica, uno de cuyos pilares es la nueva economía de la lengua en el contexto del mundo digital, científico y cultural (Wikipedia, PERTE), señaló en el panel "Lengua, inteligencia artificial e (in)dependencia tecnológica" del CILE 2023 que:

> El valor económico sin lugar a dudas también es claro porque está creciendo de manera exponencial todo el valor económico de lo que significa el mercado de PNL. […] Por eso es un proyecto estratégico económico, que en este caso con los compañeros lideramos. Está claro que también tiene una mirada de aprovechamiento económico, sino el valor económico lo van a rentabilizar otros. (CILE, 2023)

En esta línea, Cristina Gallach Figueras (2023) explicó que uno de los objetivos en el marco de esta nueva economía de la lengua impulsada por el actual Gobierno español es interconectar los diferentes corpus que existen en español, incluidos por supuesto los de la RAE, para crear un punto estable de encuentro y poder desarrollar aplicaciones digitales para el sector privado y público. Para tal fin, España cuenta con la supercomputadora desarrollada dentro del Plan de Tecnologías del Lenguaje a cargo del Centro Nacional de Supercomputación, que además es líder de la red de supercomputación iberoamericana. Expuestas así las cosas, vemos que en este terreno de las tecnologías digitales del lenguaje se entrelazan lo político, lo económico y lo tecnológico.

3.2 La variación oral: las asistentes de voz

Por otro lado, las asistentes de voz en tanto que instrumentos tecnológicos del lenguaje -las nombramos en femenino porque la mayoría tiene nombres y voces femeninas en sus

versiones en español (Alexa de Amazon; Siri de iPhone; Cortana de Microsoft o Aura de Telefónica)-, son objetos de análisis interesantes desde un punto de vista sociolingüístico porque demuestran con nitidez el lado material del lenguaje y la importancia que tienen los aspectos paraverbales y suprasegmentales en la comunicación en general y, en especial, en la co-construcción de las identidades en la era posthumanista. Desde una perspectiva glotopolítica, esta materialidad basada en la voz da pie a tratar de qué manera las normas de la cultura escrita se intersecan con nuevas normas que emergen desde las prácticas lingüísticas de la oralidad en las que rasgos lingüísticos y suprasegmentales considerados como "normales" de la propia variedad desempeñan un papel relevante, así como el influjo que pueden tener estas asistentes digitales en cómo percibimos lo normal y lo normativo en el medio oral y qué acciones lingüísticas están (o pueden estar) normalizadas y, por último, a quiénes conceptualizan los usuarios/as como agentes lingüísticos (Schneider, 2021).

Debido a que la interacción con las asistentes de voz es oral, un aspecto fundamental es que la máquina sea capaz de entender a los usuarios/as y que estos, a su vez, perciban las voces de las asistentes como "naturales". Por ello resulta fundamental con qué tipo de bases de datos orales se entrenan estas voces. La importancia de la conexión "identitaria" con dichas voces digitales se muestra de forma nítida en un spot humorístico que creó un grupo de comediantes mexicanos y donde proponen el uso de "Yatsiri", "una app que entiende a la perfección la jerga mexicana" (Mulato, 2016) precisamente por la frustración que experimenta el/la hablante de español mexicano cuando la asistente de voz le dice: "Disculpa no te entendí" – "Repítelo en español por favor". En la grabación se puede ver a cinco personas mexicanas de diferentes regiones del país haciéndole preguntas a Yatsiri en español chilango, yucateco, norteño, costeño y en lengua Tsotsil. La compañía creadora de Yatsiri, la Pina Apple Company, publicita al final del spot con lo siguiente:

Figura 1

Captura de pantalla del spot Yatsiri publicado en Facebook

Una parodia semejante se encuentra en YouTube para un asistente de voz con acento murciano, Antoñico, lo cual muestra la interrelación entre el aspecto material y corporeizado del lenguaje y las identidades. En efecto, "la lengua española asume valores materiales y simbólicos propios" (del Valle 2007, p.13) en los diferentes espacios en la que se habla y que los hablantes saben negociar en sus prácticas diarias. Pero, además, el mutuo reconocimiento de voz (por parte de la máquina y de la persona que interactúa con esta) conlleva otros aspectos relacionados con la agencia ciudadana, como exponíamos anteriormente (cfr. sección 3, *supra*).

Ahora bien, como ya mostramos también, en el mundo la tecnología digital comercial la concepción y desarrollo de tecnologías digitales del lenguaje están ligados al atractivo económico que tienen las lenguas. En otras palabras, revelan los valores del mercado y cómo se distribuye el poder social de estas lenguas. En el caso de tecnologías de voz, este atractivo se hace extensivo a las variedades espaciales, muchas de las cuales, en el caso del español, no solo constituyen formas diatópicas distintas, sino normas nacionales. Por tal motivo vale la pena analizar qué "españoles" hablan las asistentes de voz comerciales. Así, por ejemplo, la asistente de voz de Apple, que es la más conocida y antigua, ya que se creó en 2011, ofrece las siguientes opciones para el español: Chile, España, Estados Unidos y México; Alexa de Amazon y Cortana de Microsoft solo ofrece dos variedades: España y México. El asistente de Google es el que de lejos más variedades ofrece: España, Argentina, Chile, Colombia y Perú.

Si analizamos estos datos, parece bastante claro que un criterio fundamental es el número de hablantes (México, Colombia), pero también la importancia de variedades percibidas como "ejemplares" en el mundo en general, de ahí que todas estas tecnológicas ofrezcan una versión en español europeo. El caso de Estados Unidos es bastante interesante porque demuestra el empuje de este mercado hispanohablante debido al número de hablantes[10], a pesar de que el español no sea lengua oficial ni tampoco goce de prestigio abierto en ese país. Otro criterio que parece estar detrás del desarrollo de tecnologías de voz es la idiosincrasia de la variedad con respecto a otras, lo que se aplica a la variedad chilena y argentina, con rasgos bastante diferenciados de acuerdo a las percepciones de los propios hispanohablantes (Amorós Negre y Quesada Pachecho, 2019; Cestero y Paredes, 2018).

CONCLUSIONES

A partir de un análisis que combina la perspectiva glotopolítica y la posthumanista, en este artículo abro un espacio de reflexión crítico en relación a los retos que conllevan las nuevas tecnologías digitales del lenguaje para los hablantes del español y de lenguas amerindias. Me ocupo sobre todo de su papel en la (re)creación de jerarquías y sesgos

lingüísticos. Empleo la palabra *sesgo*, tal y como sugiere Blodgett et al. (2020), en la medida en que esa desviación implica una desventaja social, política o económica de una lengua/variedad (y por ende de sus hablantes) con respecto a otra(s). Teniendo esto en cuenta, hemos visto que, efectivamente, las nuevas tecnologías digitales comportan determinados sesgos debido a aspectos técnicos relacionados con la lengua fuente desde donde se crean y se desarrollan estas tecnologías, el inglés. En concreto, las técnicas de generalización y escalaridad de los procesos algorítmicos acarrean homogenización y simplificación de la riqueza léxica y de las tradiciones discursivas de la lengua española, imponiendo patrones del inglés; las cantidades de datos que procesan estas máquinas ocasionan también la regularización de procesos gramaticales en lenguas con mayor riqueza gramatical.

En casos de lenguas como las amerindias, que poseen sistemas gramaticales totalmente diferentes al inglés o al español, los programas que podrían emplearse para crear herramientas digitales del lenguaje no funcionan; producir sus propios recursos implican fuertes inversiones de recursos financieros y tecnológicos que los mercados no están dispuestos proveer. Como consecuencia, los hablantes de lenguas amerindias se ven forzados a emplear estas tecnologías digitales en español (o en otras lenguas), intensificándose de esta manera el valor de lenguas súpercentrales (De Swaan, 2013).

Otros sesgos están directamente relacionados con las jerarquías lingüísticas en las que el orden impuesto por procesos coloniales se refleja en las escasas posibilidades de creación y generación de datos en lenguas minorizadas, como por ejemplo, las lenguas amerindias, que durante siglos han sido de producción únicamente oral, reservándose la escritura a las lenguas de mayor prestigio, las coloniales. Este hecho acarrea que los hablantes de estas lenguas tengan que emplear estos recursos tecnológicos en las lenguas mayoritarias, que son las de impronta colonial. Como consecuencia de ello, siguen reproduciéndose dependencias sociolingüísticas, si tenemos en cuenta que las lenguas son un componente importante de las identidades. Las consecuencias negativas también son glotopolíticas, pues estas desventajas en el manejo de herramientas que dan acceso al mundo del siglo XXI contribuyen a erosionar sociopolítica y económicamente estas lenguas ya muy minorizadas.

Pasando ahora a lo que ocurre entre las variedades del español, vemos que, efectivamente, el valor estratégico y económico que tienen estas tecnologías digitales del lenguaje debido a la fuerza de divulgación global de prácticas semióticas a partir de sus dispositivos da pie a la creación de alianzas entre grandes empresas tecnológicas, Estados nacionales e instituciones de la lengua en la búsqueda de participar de los beneficios económicos y políticos que se generan.

Por último, los ejercicios de análisis de las herramientas de traducción comercial y de las asistentes de voz arrojan resultados bastante interesantes en relación a la jerarquía variacional con respecto al español. En el caso de las herramientas de traducción automática, lo que parece quedar claro es que no parten de las mismas bases de datos y que algunas herramientas, como DeepL, están bastante más avanzadas que otras (Google Translate y Microsoft Tanslator). En ninguna de estas se refleja un patrón que permita concluir qué norma del español prima o se privilegia; al contrario, parece ser más bien que los procesos de automatización no han sido curados o supervisados del todo. Con excepción tal vez de Deepl, lo que sí tienen en común los traductores automáticos de Google y de Microsoft es la fuerte tendencia hacia la simplificación de la compleja realidad variacional del español, muy probablemente condicionada por los propios programas empleados en el PNL. Y esto tiene repercusiones glotopolíticas nada desdeñables en la formación de las subjetividades de los hablantes. En concreto, muchas de estas herramientas de traducción automática se suelen emplear sin mayor reflexión en las escuelas (de lengua), las universidades y en las empresas; si se parte de las posibilidades léxicas que ofrecen estas herramientas, se corre el peligro de que acervos léxicos que conforman normas nacionales queden borrados.

Por su parte, el análisis de las asistentes de voz muestra que los criterios de elección para configurar y ofrecer en estos dispositivos diferentes normas del español no refleja únicamente la tradición glotopolítica de una lengua marcada durante siglos por una ideología lingüística monocéntrica, sino otros aspectos que tienen que ver con el número de hablantes (México y Colombia), el empuje de los mercados (Estados Unidos) y el conocimiento de que existen normas nacionales con aspectos muy idiosincrásicos que las alejan de aquellas más extendidas, como la mexicana o la española, por ejemplo (Argentina y Chile) y para las que, no solo para asegurar la comprensibilidad entre la máquina y los hablantes, sino reforzar el aspecto identitario de los consumidores, las grandes tecnológicas han considerado oportuno desarrollar productos diferenciados. Esto hechos demuestran con nitidez que las lenguas, en tanto productos semióticos, son fuertemente materiales y se corporeizan en las personas que las hablan.

Como ya señalamos en la introducción, este artículo debe entenderse como un espigueo sobre temas de relevancia glotopolítica en relación con las nuevas tecnologías digitales del lenguaje en español y en lenguas amerindias, acerca de las cuales aún hay mucho por hacer.

REFERENCIAS BIBLIOGRÁFICAS

Alvarado, G. (2022). ¿Y qué es ser mapuche? ¿Y qué es ser winka?" Despliegues y negociaciones de identidad en espacios digitales de Facebook dedicados a la

revitalización del mapudungun en Chile: un análisis glotopolítico. *Caracol* (24), 38 -75.

Alva, C., & Oncevay, A. (2017). Spell-Checking based on Syllabification and Character-level Graphs for a Peruvian Agglutinative Language. *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, 109-116. https://doi.org/10.18653/v1/W17-4116

Amorós Negre, C., & Quesada Pacheco, M. Á. (2019). Percepción lingüística y pluricentrismo: Análisis del binomio a la luz de los resultados del Proyecto Linguistic Identity and Attitudes in Spanish-speaking Latin America (LIAS). *ELUA*, *33*, 9. https://doi.org/10.14198/ELUA2019.33.1

Arnoux, E. N. D. (2014). Glotopolítica: delimitaciones del campo y discusiones actuales con particular referencia a Sudamérica. En L. Zajícová, R. Zámec (eds.), *Lengua y política en América Latina: Perspectivas actuales* (pp. 19-43). Univerzita Palackého v Olomouci.

Arnoux, E. N. D. (2016). La perspectiva glotopolítica en e estudio de los instrumentos lingüísticos: aspectos teóricos y metodológicos. *Matranga. Estudos Linguísticos & Literários*, *23* (38). https://doi.org/10.12957/matraga.2016.20196

Arnoux, E. N. D. (2020). De la "unidad en la diversidad" al "español auxiliar internacional". En S. Greußlich y F. Lebsanft (eds.), *El español lengua pluricéntrica. Discurso, gramática, léxico y medios de comunicación masiva* (pp. 39-60). Bonn University Press.

Arnoux, E. N. D. & del Valle, J. (2010). La representaciones ideológicas del lenguaje. Discurso glotopolítico y panhispanismo. *Spanish in Context*, *7*(1), 1-24.

Arnoux, E. N. D., & Lauria, D. (2023). La prescripción en los discursos sobre la lengua. En C. López Ferrero, I. E. Carranza, & T. A. Van Dijk, *Estudios del discurso* (1.ª ed., pp. 129-142). Routledge. https://doi.org/10.4324/9780367810214-12

Bein, R. (2020). Los desafíos de una ciencia plurilingüe (también en tiempos de pandemia). En F. Dandrea y G. Lizabe (eds.), *Internacionalización y gobernanza lingüística en el nivel superior: las lenguas extranjeras en contexto* (pp. 13– 28). UniRío Editora.

Bertolotti, V. (2023). *En Lengua, inteligencia artificial e (in)dependencia tecnológica #CILE2023* [Video]. YouTube. https://www.youtube.com/watch?v=mWmVb2BtB8Y&t=1503s (min. 36.30)

Blodgett, S. L., Barocas, S., Daumé Iii, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454-5476. https://doi.org/10.18653/v1/2020.acl-main.485

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). *On the Opportunities and Risks of Foundation Models*. Stanford University https://doi.org/10.48550/ARXIV.2108.07258

Bourdieu, P. (1985). *¿Qué significa hablar?* Akal.

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5), 585-614.

Bustamante, G., Oncevay, A., & Zariquiey, R. (2020). No data to crawl? Monolingual corpus creation from PDF files of truly low-resource languages in Peru. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2914-2923. European Language Resources Association.

Caravedo, R. (2014). *Percepción y variación lingüística: Enfoque sociocognitivo*. Iberoamericana; Vervuert.

Cestero, A. M., & Paredes, F. (2018). Creencias y actitudes hacia las variedades cultas del español actual: El proyecto PRECAVES XXI. *Boletín de Filología*, *53*(2), 11-43. https://doi.org/10.4067/S0718-93032018000200011

Chomsky, N., Roberts, I., & Watumull, J. (2023, marzo 8). Noam Chomsky: The False Promise of ChatGPT. *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Coronel-Molina, S. M. (2019). Media and Technology: Revitalizing Latin American Indigenous Languages in Cyberspace. En T.L. McCarty, S.E. Nicholas y G. Wigglesworth (eds.), *A world of Indigenous languages: Politics, pedagogies and prospects for language reclamation* (pp. 91-114). Multilingual Matters.

Cuantum Technologies (2023). *Introduction to natural language processing with Transformers*. Cuantum Technologies.

Danet, B., & Herring, S.C. (2003). Introduction: The Multilingual Internet. *Journal of Computer-Mediated Communication*, *9*(1). https://doi.org/10.1111/j.1083-6101.2003.tb00354.x

De Swaan, A. (2013). *Words of the World: The Global Language System*. Wiley.

Del Valle, J. (ed.) (2007). *La lengua, ¿patria común?: ideas e ideologías del español*. Vervuert/Iberoamericana.

Del Valle, J. (2007). La lengua, patria común: la hispanofonía y el nacionalismo panhispánico. En J. D. Valle (ed.) *La lengua, ¿patria común?: ideas e ideologías del español* (pp. 31-56). Vervuert/Iberoamericana.

Del Valle, J. (2017). La perspectiva glotopolítica y la normatividad. *Anuario de Glotopolítica*, *1*, 17-40.

Fergus, P., & Chalmers, C. (2022). Natural Language Processing. En P. Fergus & C. Chalmers, *Applied Deep Learning* (pp. 217-244). Springer International Publishing. https://doi.org/10.1007/978-3-031-04420-5_9

Ferrando, F. & Brito Ledesma, J. I. (2022). Posthumanismo, Transhumanismo, Antihumanismo, Metahumanismo y Nuevos Materialismos: Diferencias y Relaciones. *Revista Ethika+*, *5*, 151-166. https://doi.org/10.5354/2452-6037.2022.65842

Gallach, C. (2023). Inversión en tecnología e inteligencia artificial. En *Lengua, inteligencia artificial e (in)dependencia tecnológica #CILE2023*. [Video]. YouTube. https://www.youtube.com/watch?v=mWmVb2BtB8Y&t=1503s (min. 57.09)

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P. & Villegas, M. (2021). Maria: Spanish language models. Arxiv preprint. *arXiv:2107.07253*.

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2021). *GlottoScope*. Glottolog.

Haas, W. (Ed). (1982). *Standard languages: spoken and written*. University Press.

Heyd, T., & Schneider, B. (2019). The sociolinguistics of late modern publics. *Journal of Sociolinguistics*, *23*(5), 435-449. https://doi.org/10.1111/josl.12378

Instituto Cervantes. (2023, marzo 30). *Lengua, inteligencia artificial e (in)dependencia tecnológica. #CILE2023*. [Video] YouTube. https://www.youtube.com/watch?v=mWmVb2BtB8Y&t=3974s&ab_channel=InstitutoCervantes

Instituto Cervantes (2023). *El español en el mundo. Anuario del Instituto Cervantes.* En Centro Virtual Cervantes. https://cvc.cervantes.es/lengua/anuario/anuario_23/

Internet World Users by Language. (2020). *Internet World Stats*. https://www.internetworldstats.com/stats7.htm

Kelly-Holmes, H. (2019). Multilingualism and Technology: A Review of Developments in Digital Communication from Monolingualism to Idiolingualism. *Annual Review of Applied Linguistics*, *39*, 24-39. https://doi.org/10.1017/S0267190519000102

Lara, L. F. (2007). Por una reconstrucción de la idea de la lengua española. En J. D. Valle (ed). *La lengua, ¿patria común?* (pp. 163-181). Vervuert/Iberoamericana.

Lara, L. F. (2015). *Temas del español contemporáneo*. El Colegio de México.

Moreno, A. (s. f.). Procesamiento del lenguaje natural ¿qué es? *Instituto de ingeniería del conocimiento*. https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/

Moreno Sandoval, A. (2019). *Lenguas y computación*. Síntesis.

Mulato, A. (2016, octubre 25). *Esta versión de Siri sí entiende a los mexicanos*. Verne; Ediciones El País. https://verne.elpais.com/verne/2016/10/25/mexico/1477429578_365876.html

Muñoz Machado, S. (2021, junio 21). *Encuentro sobre Lengua Española e Inteligencia Artificial*. [Video] YouTube. https://www.youtube.com/watch?v=A4JL9qCuylo&t=1356s&ab_channel=RAEInforma (min. 07.18)

Navarro, F., Lillis, T., Donahue, T., et al. (2022). Rethinking English as a lingua franca in scientific-academic contexts: A position statement. *Journal of English for Research Publication Purposes*, *3*(1), 143-153. https://doi.org/10.1075/jerpp.21012.nav

Oncevay, A., Cardoso, G., Alva, C., et al. (2022). SchAman: Spell-Checking Resources and Benchmark for Endangered Languages from Amazonia. En Y. He, H. Ji, et al. (Eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022—Volume 2: Short Papers, Online only, November 20-23, 2022* (pp. 411-417). Association for Computational Linguistics. https://aclanthology.org/2022.aacl-short.51

Ortega, J. E., Castro-Mamani, R. A., & Montoya Samame, J. R. (2020). Overcoming Resistance: The Normalization of an Amazonian Tribal Language. *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, 1-13.

Passban, P., Way, A., & Liu, Q. (2018). Tailoring Neural Architectures for Translating from Morphologically Rich Languages. En *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3134-3145). Association for Computational Linguistics.

Pennycook, A. (2016). Posthumanist Applied Linguistics. *Applied Linguistics*, *39*(4), 445-461. *https://doi.org/10.1093/applin/amw016*

Real Academia Española (en línea). *Corpus de Referencia del español Actual (CREA).* Versión anotada. https://apps2.rae.es/CREA/view/inicioExterno.view

Real Academia Española (en línea). *Corpus del Español del Siglo XXI (CORPES).* http://www.rae.es

Ravindran, S. (2023). Lost in translation. *Science*, *381*(6655), 262-265. https://doi.org/10.1126/science.adj8519

Reich, R., Sahami, M., & Weinstein, J. M. (2021). *System error: Where big tech went wrong and how we can reboot* (First edition). Harper.

Schneider, B. (2021). Von Gutenberg zu Alexa: Posthumanistische Perspektiven auf Sprachideologie. En M. Lind (Ed.), *Mensch—Tier—Maschine* (pp. 327-346). transcript Verlag. https://doi.org/10.1515/9783839453131-014

Schneider, B. (2022). Multilingualism and AI: The Regimentation of Language in the Age of Digital Capitalism. *Signs and Society*, *10*(3), 362-387. https://doi.org/10.1086/721757

Sennrich, R. (2023, abril 31). *Interaktive Sprachmodelle: Ein Blick hinter die Kulissen*. Interaktive Sprachmodelle: Lehre und Forschung mit ChatGPT & Co, Bern. https://www.hd.unibe.ch/kurse___tagungen/kilof_tagung/index_ger.html

Sullón Acosta, K. N. (Ed.). (2013). *Lenguas originarias del Perú*. Ministerio de Educación.

Telefónica. (s. f.-a). *Estrategia*. https://www.telefonica.com/es/nosotros/estrategia/

Telefónica. (s. f.-b). *Países y Unidades emergentes*. https://www.telefonica.com/es/nosotros/paises-unidades-emergentes/

The decoder (2023, julio 11). GPT-4 architecture, datasets, costs and more leaked. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/#:~:text=Training%20Cost%3A%20The%20training%20costs%20for%20GPT-4%20was,the%20larger%20clusters%20required%20and%20lower%20utilization%20rates.

United States Census Bureau. (2023, agosto 15). *How has the racial and ethnic makeup of the US changed*? https://usafacts.org/data/topics/people-society/population-and-demographics/our-changing-population/

Vanmassenhove, E., Shterionov, D., & Gwilliam, M. (2021). Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2203-2213. https://arxiv.org/abs/2102.00287

Vanmassenhove, E., Shterionov, D., & Way, A. (2019). Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. *Proceedings of Machine Translation Summit XVII: Research Track*, 222-232. https://arxiv.org/abs/1906.12068

videos de humor y más 😂😂 funny videos (2022, junio 15). *Primer asistente de voz murciano*. YouTube. https://www.youtube.com/watch?v=U8cqP190aec&t=3s&ab_channel=videosdehumorym%C3%A1s%F0%9F%98%82%F0%9F%98%82funnyvideos

Wikipedia, perte (2023, Agosto 09) https://es.wikipedia.org/wiki/PERTE_(2023

Zariquiey, R. (2023). Lengua, inteligencia artificial e (i)dependencia tecnológica: reflexiones desde los idiomas amerindios. *En Lengua, inteligencia artificial e (in)dependencia tecnológica #CILE2023*. YouTube. https://www.youtube.com/watch?v=mWmVb2BtB8Y&t=1503s (min. 36.30)

[1] Yvette Bürki es profesora titular en el Instituto de Lengua y Literaturas Hispánicas de la Universidad de Berna (Suiza), donde dirige la sección de Lingüística. Sus campos de investigación se centran en la sociolingüística crítica, el análisis crítico del discurso y la lingüística sociocultural. Desde estas perspectivas, estudia las variedades periféricas del español, en especial del ladino o judeoespañol, así como manifestaciones semióticas discursivas con respecto a grupos étnico-lingüísticos minorizados, también en contextos de migración, con foco en las identidades e ideologías lingüísticas. Además, y frente a estos ejes de investigación bien establecidos, está investigando actualmente desde una perspectiva crítica las consecuencias sociolingüísticas y glotopolíticas, en particular, de las tecnologías digitales del lenguaje.

[2] De acuerdo con el informe de Juniper Research (https://www.juniperresearch.com/press/number-of-voice-assistant-devices-in-use), en 2024 habrá más de 8.400 millones de asistentes de voz, el doble que en 2020. Estos asistentes de voz se están integrando de manera progresiva no solo en el ámbito empresarial, sino también al doméstico.

[3] Los programas para el análisis basado en IA utilizan la potencia de algoritmos de aprendizaje automático y lenguajes de procesamiento natural para comprender el contexto en el que se utilizan determinadas palabras y frases e interpretar, de esta manera, la mayoría de las emociones ocultas en un texto.

[4] MarIA incluye los modelos del lenguaje en español RoBERTa-base, RoBERTa-large, GPT2 y GPT2-large que pueden considerarse como los modelos más grandes y más eficientes para español. Los modelos han sido preentrenados utilizando un corpus masivo de 570GB de textos limpios y de duplicados, que comprende un total de 135 mil millones de palabras extraídas del Archivo Web del Español construido por la Biblioteca Nacional de España entre los años 2009 y 2019 (Gutiérrez-Fandiño et al., 2022).

[5] El Instituto de Ingeniería y Ciencias del conocimiento se creó en 1989 gracias a un convenio entre IBM España y la Universidad Autónoma de Madrid (UAM), con la colaboración de la Dirección General de Informática y Nuevas Tecnologías del Ministerio de Industria y Energía con los objetivos de difusión y transferencia tecnológica (https://www.iic.uam.es/iic/historia/).

[6] A este respecto anotaré que otras herramientas de tecnología de lengua más avanzadas y globales como Bing de Microsoft y Google Bard sí dan diferentes opciones para la traducción de swimming pool si se les pregunta. Copio aquí mi interacción con Bing el 30-08-2023:

Y: “How do you say “I went to the swimming pool” in Spanish?

B: “I went to the swimming pool” in Spanish is “Fui a la piscina”

Y: Thank you. Is there another way to say the same phase in Spanish for other Spanish-Speaking countries?

B: Yes, there are other ways to say “I went to the swimming pool in Spanish depending on the country. For example, in Mexico you can say “fui a nada en la alberca”. In River Plate, “you can say fui a nadar en la pileta”.

[7] La RAE y las ASALE no emplean el término pluricentrismo, sino policentrismo. Esta diferencia de prefijo es glotopolíticamente muy significativa, ya que se inserta en el engranaje de la nueva política lingüística panhispánica, aplicada desde 2004, y que se refleja en el lema “Unidad en la diversidad”. A diferencia del pluricentrismo, que pone solo énfasis en la existencia de diferentes centros normativos de una lengua, el policentrismo, de acuerdo con la RAE y la ASALE, implica la existencia de diferentes normas regionales, ante las cuales se superpone una “supernorma” como un producto de todas las normas regionales comunes. Tal solución, según la RAE, “no solo no pone en peligro la unidad del español, sino que contribuye más bien a fortalecerla, y ayuda a comprender su distribución geográfica de forma más cabal” (Nueva Gramática de la Lengua Española, 2009, p.XXXIX-XLVIII).

[8] Por ejemplo, el Corpus de Referencia del Español Actual (CREA) está compuesto por 50% de material procedente de España y el otro 50%, de América, pero no todos los países tienen igual peso, ya que el parámetro elegido ha sido regional o zonal y no nacional. Además, de manera relacional, estos porcentajes resultan bastante desequilibrados, dado que de los aproximadamente 599 millones de hablantes que viven en países hispanohablantes (El español en el mundo. Anuario del Instituto Cervantes, 2023), aproximadamente 551 corresponden a los hablantes del español en el continente americano. En cuanto al medio, 90% corresponde a la lengua escrita, con una marcada presencia del material procedente de libros, y solo 10% a la lengua oral (https://www.rae.es/banco-de-datos/crea/parametros-habituales-crea). Sobre los problemas no solo metódologicos, sino glotopolíticos del corpus CREA véase Lara (2007, p.177, n. 25). El Corpus del español del Siglo XXI (CORPES) es aún un recurso en construcción "con desequilibrios y ajustes que irán desapareciendo en las versiones posteriores" (https://www.rae.es/corpes/), pero en líneas generales, es un corpus mucho mejor equilibrado, ya que asigna "algo más del 30%" (https://www.rae.es/corpes/contenidos/datos) del total de las formas del corpus a España y el resto a América, de acuerdo a las zonas/regiones lingüísticas ya explicitadas para el CREA.

[9] Sobre la interrelación entre la actual política panhispánica y el valor económico del español para las grandes corporaciones transnacionales de hechura española como Telefónica, BBVA; Santander; Iberdrola, etc., véanse del Valle, (2007); Lara, (2015) y Arnoux, (2020). Según el informe de la propia Telefónica de 2017 (presentado por la Fundación Telefónica el 13/02/2017), las actividades relacionadas con el idioma español generan 16% del valor económico de Producto Interior Bruto (citado en Arnoux 2020).

[10] De acuerdo a las estimaciones de las cifras poblacionales que proporciona la United States Census Bureau para el año 2022, la población hispana o latina asciende a 63.553.639 de un total de 333.287.557 habitantes. Esta cifra representa 19.1% del total de la población estadounidense. Un dato aún más interesante con respecto al empuje al que me he referido arriba es que la población latina o hispana es la que más ha crecido en los últimos cuatro años (2.7%), mientras que la población blanca (no hispana) es la que más ha disminuido (4.9%).