

INTELIGENCIA ARTIFICIAL, TECNOLOGÍAS Y RECURSOS DEL LENGUAJE: POLÍTICAS Y DERECHO PARA LA EXPLOTACIÓN DE CORPUS Y BASES DE DATOS

Lorenzo Cotino Hueso*

Resumen

El procesamiento automático del lenguaje natural y, en particular, la traducción automática tienen enorme potencial para el español y otras lenguas españolas. Los poderes públicos desde hace una década han fijado los objetivos de las políticas en inteligencia artificial (IA) y lenguas, dotados ahora con más de 1.100 millones de euros en los Proyectos Estratégicos para la Recuperación y Transformación Económica (PERTE) “Nueva Economía de la Lengua”. Es clave la generación (pública) de infraestructuras, recursos, bases y, sobre todo, corpus lingüísticos que *alimentan* a la IA y otras tecnologías lingüísticas desarrolladas especialmente por el sector privado. El Derecho de la UE tiende hacia la flexibilidad y apertura de estos recursos lingüísticos. No obstante, desde el régimen de propiedad intelectual y el de reutilización de datos, no hay obligación de abrir y poner a disposición estos recursos, sino que los promotores tienen el derecho *sui generis* a que no se pueda hacer minería de datos u otros procesados sin su autorización. Se critica la insuficiente regulación. Sobre esta base, es clave la elección de licencias permisivas, como las del ecosistema Meta-Share, de potencial proyección a los corpus españoles. Finalmente, se exponen los elementos clave en la planificación y adopción de modelos de explotación y sostenibilidad de los recursos lingüísticos en España.


Palabras clave: inteligencia artificial; tecnologías de la lengua; bases de datos lingüísticas; propiedad intelectual; derecho.

ARTIFICIAL INTELLIGENCE, LANGUAGE TECHNOLOGIES AND RESOURCES: POLICIES AND LEGISLATION FOR THE USE OF CORPORA AND DATABASES.

Abstract

*Automatic natural language processing and, in particular, machine translation offer an enormous potential for Spanish and other languages spoken in Spain. For the last decade, public authorities have set policy objectives in artificial intelligence (AI) and languages, which are now endowed with over 1.1 billion euros in the Strategic Projects for Economic Recovery and Transformation (PERTE) “New Language Economy”. A key factor is the (public) generation of infrastructures, resources, databases and, above all, language corpora that feed AI and other language technologies developed particularly by the private sector. EU Law tends towards flexibility and open use of these language resources. However, under intellectual property and data reuse regulations, there is no obligation to open these resources and make them available. On the other hand, the developers have the *sui generis* right to not allow data mining or other processing without their authorisation. Regulation is criticised for being insufficient. Given this situation, it is essential to choose permissive licences, such as those of the Meta-Share ecosystem, which has potential for extrapolation to Spanish corpora. Lastly, the key elements for planning and adopting models for language resource use and sustainability in Spain are discussed.*

Keywords: artificial intelligence; language technologies; language databases; intellectual property; law.

* Lorenzo Cotino Hueso, investigador de la Universidad Católica de Colombia y catedrático de Derecho Constitucional en la Universidad de Valencia. cotino@uv.es.  0000-0003-2661-0010.

Artículo recibido el 28.08.2022. Evaluaciones ciegas: 16.02.2023 y 17.02.2023. Aceptación de la versión final: 22.03.2023.

Citación recomendada: Cotino Hueso, Lorenzo. (2023). Inteligencia artificial, tecnologías y recursos del lenguaje: políticas y derecho para la explotación de corpus y bases de datos. *Revista de Llengua i Dret, Journal of Language and Law*, 79, 61-77. <https://doi.org/10.58992/rld.i79.2023.3860>

Sumario

- 1 Los textos como *big data* que “alimenta” a la inteligencia artificial
 - 2 Tecnologías del lenguaje, minería de datos y una inteligencia artificial que “piense en español”
 - 3 Infraestructuras y recursos lingüísticos, corpus y bases de conocimiento: elementos esenciales de las políticas digitales de la lengua en España
 - 4 Los corpus lingüísticos de referencia en España
 - 5 Políticas de tecnologías y recursos lingüísticos y el papel del sector público
 - 6 La concurrencia y variabilidad de la regulación aplicable a los recursos lingüísticos
 - 7 Los titulares de los recursos lingüísticos tienen el derecho sui generis a que no se pueda hacer minería de datos sin su autorización, salvo excepciones
 - 8 Las exigencias de la regulación de reutilización y la interoperabilidad de los recursos lingüísticos
 - 9 Hacia las licencias permisivas de la explotación de recursos lingüísticos: las licencias permisivas del ecosistema del Meta-Share y su proyección a los corpus españoles
 - 10 Para concluir: la necesidad de planes y modelos de explotación y sostenibilidad de los recursos lingüísticos en España
- Referencias

* 1 Los textos como *big data* que “alimenta” a la inteligencia artificial

La minería de datos, algoritmos, sistemas de inteligencia artificial (IA) con aprendizaje automático o profundo (*machine learning*, *deep learning*), robótica y otras tecnologías convergentes “beben” del *big data* o de los macrodatos. Se habla de las “V”: volumen, variedad, velocidad y valor, a las que se añaden entre otras, la veracidad (Cotino, 2020a). La noción de *big data* o macrodatos integra, asimismo, el propio tratamiento masivo de los datos. La Resolución del Parlamento Europeo, de 14 de marzo de 2017, sobre macrodatos¹ afirma, en la letra A, que “el concepto de macrodato se refiere a la recopilación, análisis y acumulación constante de grandes cantidades de datos, incluidos datos personales, procedentes de diferentes fuentes y objeto de un tratamiento automatizado mediante algoritmos informáticos y avanzadas técnicas de tratamiento de datos, utilizando tanto datos almacenados como datos transmitidos en flujo continuo, con el fin de generar correlaciones, tendencias y patrones (analítica de macrodatos)”.

Los datos pueden ser estructurados, semiestructurados o no estructurados (Ortega Giménez, 2019, pp. 176-178); los datos y fuentes no estructurados son muchos más y más variados (Alcalde Bezhold y Alfonso Farnós, 2019, pp. 60-61). Se trata de datos que no tienen modelo u organización. Pueden ser los datos generados por el uso de redes, aplicaciones, sensores, sistemas máquina a máquina o grandes transacciones de gestión de atención y facturación, biometría, etc.

Los textos de múltiples procedencias son un *alimento* de la IA de extraordinario valor. En el contexto de la lengua, se afirma que más del 90 % de la información digital disponible es información no estructurada en forma de textos y documentos (escritos o hablados) en múltiples lenguas (Secretaría de Estado de Telecomunicaciones y Sociedad de la Información [SETSI], 2015a, p. 6). Así, las infraestructuras, recursos lingüísticos, corpus y bases de datos se nutren, en buena medida, de datos nada o escasamente estructurados. Y las tecnologías del lenguaje con la IA a la cabeza son esenciales para extraer su valor.

Tras el inglés, el chino y, según los casos, el hindi, el español es una de las lenguas con más hablantes nativos del mundo (474,7 millones) y más usuarios (548,3 millones), la tercera en usuarios de la red y, también, entre las primeras por intercambio económico (Wikipedia, 2022; Ministerio de Asuntos Exteriores, 2022). Nueve de cada diez hablantes están fuera de España, y se prevé que, para 2060, Estados Unidos sea el segundo país hispanohablante del mundo (Ministerio de Asuntos Económicos y Transformación Digital [MAETD], 2022a, p. 5). La relevancia estratégica de esta lengua y su vinculación con las tecnologías disruptivas no ha pasado desapercibida en las políticas en España. En cualquier caso, estas políticas y planes no descuidan otras lenguas oficiales a las que Tasa Fuster (2022) ha prestado especial atención. Bien es cierto, que, como afirma Rigau (2022, p. 201): “Existe una distancia abismal entre ellas. En particular, en el desarrollo de la TL en castellano frente al inglés, y en el desarrollo de la TL en catalán, gallego y euskera con respecto al castellano, y por supuesto al inglés”.

2 Tecnologías del lenguaje, minería de datos y una inteligencia artificial que “piense en español”

El lenguaje es una de las herramientas distintivas del ser humano, y las matemáticas son el soporte de nuestra ciencia y base de las TIC. Pues bien, lengua y matemáticas se aúnan en “herramientas”, “tecnologías”, “sistemas” o “aplicaciones” “lingüísticas”, lo que podemos denominar como *tecnologías del procesamiento del lenguaje natural y de la traducción automática* (SETSI, 2015b, pp. 6 y 37). Estas tecnologías lingüísticas están en el *corazón* del *software* que actualmente procesa la información no estructurada y explota la gran cantidad de datos contenidos en los textos, también los de la web, plataformas y medios sociales (SETSI, 2015b, pp. 6 y 37; Secretaría de Estado para la Sociedad de la Información y la Agenda Digital [SESIAD], 2018, p. 19 y ss.). Las aplicaciones en la industria del lenguaje son muy variadas. Por una parte, herramientas y motores

* El presente estudio es resultado de investigación del proyecto “Derecho, Cambio Climático y Big Data”, Universidad Católica de Colombia; MICINN: Retos “Derechos y garantías frente a las decisiones automatizadas...” (RTI2018-097172-B-C21 y PID2022-136439OB-I00); “La regulación de la transformación digital ...” grupo de investigación de excelencia Generalitat Valenciana “Algorithmic law” (Prometeo/2021/009, 2021-24); “Algorithmic Decisions and the Law: Opening the Black Box” (TED2021-131472A-I00); “Transición digital de las Administraciones públicas e inteligencia artificial” (TED2021-132191B-I00) del Plan de Recuperación, Transformación y Resiliencia. Estancia Generalitat Valenciana CIAEST/2022/1.

¹ [Resolución del Parlamento Europeo, de 14 de marzo de 2017, sobre las implicaciones de los macrodatos en los derechos fundamentales: privacidad, protección de datos, no discriminación, seguridad y aplicación de la ley](#) (DOUE C, núm. 263, 25.07.2018, pp. 82-89).

de traducción automática y traducción asistida por ordenador. Por otra parte, el conjunto de tecnologías de procesamiento del habla y sistemas conversacionales, que incluye los sistemas de reconocimiento del habla, la comprensión del lenguaje hablado, la conversión a texto, la síntesis de voz a partir de texto, la interacción, negociación y generación del habla, los sistemas de diálogo o asistentes conversacionales, los sistemas de conversión de habla a texto, y viceversa (*speech-to-text* y *text-to-speech*, respectivamente), así como sistemas que realizan tareas de preprocesamiento morfosintácticas, sintácticas, semánticas, pragmáticas y discursivas. Se trata de aplicaciones informáticas que sirven para la clasificación, la agrupación, el filtrado y el enrutamiento de documentos, la creación de resúmenes automáticos, la extracción de datos, el análisis de sentimientos y minería de opinión, el seguimiento y monitorización de la reputación en los medios sociales, la creación de alertas, la corrección ortográfica y gramatical, la ayuda al aprendizaje y la enseñanza de idiomas, la búsqueda inteligente y optimizada de documentos, los sistemas de respuesta automática a preguntas, los asistentes personales o la traducción automática de textos (SESIAD, 2018, pp. 18-19; SETSI, 2015b, p. 12).

Es así como los macrodatos, generalmente no estructurados, pasan a suponer información y conocimiento que puede llegar a ser muy valioso. Su uso transforma nuestra relación con los sistemas informáticos y artificiales, este *big data* favorece la creación de una sociedad multilingüe, está detrás de muchos productos y plataformas digitales. La IA se encuentra en muchos casos en el centro de estas tecnologías lingüísticas, en clara convergencia con otras tecnologías afines, como la minería de datos y *text analytics* (habitualmente con aprendizaje automático), *question answering*, *text entailment* (implicación lógica en textos) (SESIAD, 2018, p. 25-26). Bien es cierto que otras tecnologías lingüísticas no dependen de conjuntos no estructurados de datos, como pueda ser Apertium, que usa reglas y diccionarios.

Las herramientas digitales son el elemento fundamental de la economía del conocimiento, y la IA es “uno de los elementos disruptivos fundamentales” (MAETD, 2022a, p. 15). De hecho, la IA ha pasado a ocupar el lugar central y más visible tanto en los objetivos como en los medios de las políticas lingüísticas. Así, la IA está presente por cuanto se busca:

Impulsar un ecosistema de innovación de inteligencia artificial en español [...], asegurar que las nuevas tecnologías incorporan los principios de nuestras lenguas [...] y garantizar que la inteligencia artificial procese correctamente en español, mediante el impulso a toda la cadena de valor de la industria-servicios en la nueva economía de la lengua” (MAETD, 2022a, pp. 8-9).

Como se verá, se trata de generar bases de conocimiento y corpus que alimenten a la IA (eje transversal del PERTE,) y, directamente, como eje II, se busca que “la inteligencia artificial procese adecuadamente en español, que ‘piense en español’” (MAETD, 2022a, pp. 14 y ss.)

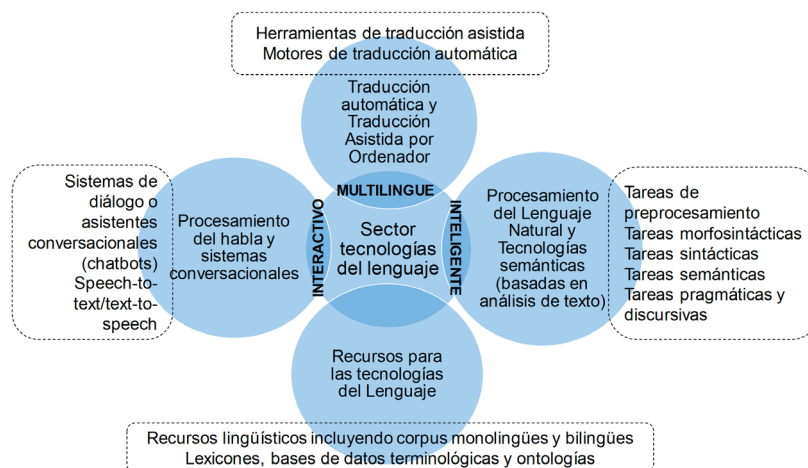
Pese a la visibilidad mayor de la IA en las políticas, desde el punto de vista normativo, en la Directiva 2019/790,² el tratamiento, uso o explotación de datos a través de tecnologías disruptivas e IA se reconduce al concepto más amplio de “minería de textos y datos” (art. 2.2).³ En el mismo caben también técnicas que no son IA, sino que están basadas en reglas o estadística.

2 Directiva 2019/790/CE del Parlamento Europeo y del Consejo, de 17 de abril de 2019, sobre los derechos de autor y derechos afines en el mercado único digital y por la que se modifican las Directivas 96/9/CE y 2001/29/CE (DOUE L, núm. 130, 17.05.2019, pp. 92-125).

3 Concepto que se define como: “Toda técnica analítica automatizada destinada a analizar textos y datos en formato digital a fin de generar información que incluye, sin carácter exhaustivo, pautas, tendencias o correlaciones”.

Gráfico 1

Clasificación de las soluciones de tecnologías del lenguaje



Fuente. SESIAD, 2018, p. 22.

3 Infraestructuras y recursos lingüísticos, corpus y bases de conocimiento: elementos esenciales de las políticas digitales de la lengua en España

Según se ha adelantado, una inmensa cantidad de datos no están estructurados. Y, en buena medida, estos datos se agrupan en lo que se denomina *infraestructura y recursos lingüísticos*. Se trata de grandes cantidades de textos en formato electrónico que sirven de ejemplo para los algoritmos de aprendizaje automático y otras tecnologías del lenguaje. Los recursos lingüísticos pueden estar en varios formatos. Si bien la base son datos no estructurados (es decir, datos en bruto), normalmente, para el procesamiento lingüístico, es necesario que pasen a ser primarios; esto es, que dejen de ser datos en bruto porque ya cuenten con algún tipo de información estructural del documento. Y se habla de *datos secundarios* cuando, además, contienen anotaciones que añaden valor a los datos originales (por ejemplo, con metadatos sobre el origen, lengua, dominio), y pueden estar anotados a diferentes niveles desde el punto de vista lingüístico (SETSI, 2015a, p. 6; SETSI, 2015b, p. 10). Se habla de *recursos derivados de otros ya existentes* cuando se han combinado datos primarios o secundarios. Así, se pasaría de corpus monolingües a corpus bilingües, o se realizaría la inducción de un léxico a partir de un corpus de textos.

Entre los recursos lingüísticos se habla de *corpus* y de *bases de datos* para hacer referencia a colecciones de datos secundarios (algo estructurados al contar con metadatos). Se emplea *corpus* para referirse a colecciones de texto escrito y *base de datos* para referirse a grabaciones de voz, a colecciones de términos, etc. (tesauros, WordNet, PropBank). Son ejemplos de recursos lingüísticos los corpus monolingües, los paralelos o memorias de traducción, los multilingües y los multimodales, los de texto analizado sintácticamente (*treebanks*, *dependency banks*), los semánticos (SemCor, PropBank), los de discurso y los de referencia. Asimismo, entre los recursos lingüísticos, cabe mencionar los léxicos, los lexicones⁴ o diccionarios, las ontologías⁵ y las gramáticas.

Pese a la importancia y potencialidad del español en el mundo, esta lengua tiene escaso protagonismo para entrenar a los algoritmos de IA, y representa menos del 30 % del mercado mundial de las tecnologías de procesamiento de lenguaje natural (MAETD, 2022a, p. 15). La situación del inglés y de otros idiomas es bastante más ventajosa en términos de cantidad, disponibilidad, calidad, cobertura, madurez, sostenibilidad

4 Los lexicones van más allá de un simple repositorio de palabras, dado que “integra[n] conocimiento fonológico (información sobre la entonación o la acentuación), morfológico (sobre la estructura de las palabras), sintáctico (sobre cómo se organizan las palabras en frases), semántico (sobre el significado de las palabras y de cómo estos significados se combinan en las oraciones) y pragmático (información sobre la intencionalidad asociada al uso del lenguaje)” (SESIAD, 2018, pp. 20-21).

5 Una ontología puede definirse como “el vocabulario común y unificado de términos, que representa adecuadamente el significado (semántica) de los conceptos más usados en un sector específico. También se consideran como ontología [...] los tesauros y modelos conceptuales” (SESIAD, 2018, p. 21).

y adaptabilidad de los recursos. Son muchas e importantes las empresas que se dedican al procesamiento del inglés y tienen a su disposición una extensa colección de procesadores y recursos lingüísticos de acceso libre o con licencias asequibles: procesadores lingüísticos (CoreNLP, OpenNLP), recursos lingüísticos y semánticos (WordNet, FrameNet o VerbNet), grandes corpus de texto en bruto (ClueWeb o GigaWord), así como corpus de gran tamaño anotados con información lingüística variada y de alta calidad (Ontonotes) (SETSI, 2015a, p. 33). Ahora bien, como se ha podido comprobar en la red, en los últimos años, estas empresas disponen de sus recursos también en español.⁶

Además de las tecnologías del lenguaje, la creación y el uso de recursos e infraestructuras lingüísticas son de gran importancia para un gran número de aplicaciones avanzadas. De ahí que las políticas públicas se centran especialmente en infraestructuras y recursos lingüísticos por cuanto aumentan el número, la calidad, la variedad y la disponibilidad de las herramientas, aplicaciones y procesadores que les dan soporte como una infraestructura común, para reducir, así, la inversión inicial que permitiría a las pymes españolas aprovecharse de esta oportunidad (SETSI, 2015b, p. 8).

La European Language Resources Association (ELRA) es la distribuidora europea de recursos y herramientas lingüísticas desde 1995.⁷ En su contexto, la [Multilingual Europe Technology Alliance \(META\)](#) es una red de excelencia creada por la Comisión Europea. Sus documentos, libros blancos y agendas han sido la clara referencia para las políticas de tecnologías y recursos lingüísticos en España y en la Unión Europea (UE). En España, desde META, la SETSI tuvo en cuenta el libro blanco *La lengua española en la era digital* (Melero et al., 2012) al elaborar el *Informe* (2015a) y, posteriormente, el *Plan* (2015b). También hubo libros blancos para el gallego, el catalán y el vasco.

La SETSI, en el *Informe* (2015a), recomendó crear una plataforma tecnológica del español para elaborar un repertorio del español científico-técnico homogéneo y dar impulso a la difusión de textos especializados en español.

Estos objetivos se aprecian desde las primeras políticas digitales lingüísticas y, con más intensidad si cabe con los impulsos de la Estrategia Nacional de Inteligencia Artificial (ENIA), de diciembre de 2020 (MAETD, 2020c), y el PERTE “Nueva Economía de la Lengua”, de marzo de 2022 (MAETD, 2022a, 2022b),⁸ en el que la generación de bases de conocimiento es el eje transversal; este eje está dotado con 96,7 millones de euros. Se recuerda, precisamente, que es la “base de la cadena de valor” y un “elemento crítico” para que “pueda ser aprovechada por todas las industrias implicadas”, “información de calidad procesada por la misma de forma sencilla y centralizada” (MAETD, 2022a, pp. 9 y ss.).

La producción de un recurso lingüístico implica diversas tareas o fases (SETSI, 2015a, p. 37) que bien pueden ser llevadas a cabo por los mismos agentes o por diferentes. Así, se requiere de la concepción inicial y la especificación del recurso lingüístico del que se trate. De ahí se pasa a la producción del recurso, tarea esencial que requiere la compilación, depuración o generación de nuevos datos a partir de datos primarios o secundarios. Asimismo, son precisas las tareas de validación, esto es, de evaluación y control de calidad del recurso. Igualmente, entre las tareas o fases, cabe mencionar la tarea de mantenimiento, con mejoras, ampliaciones y resolución de errores. Un elemento esencial es la planificación y decisión de la explotación, puesta a disposición, distribución y acceso al recurso lingüístico, corpus o bases de datos creado. Lo natural es que los recursos lingüísticos se reutilicen y se vuelvan a procesar para la generación de nuevos recursos o, especialmente, para la aplicación misma de tecnologías y aplicaciones del lenguaje. El modelo de explotación

⁶ <https://stanfordnlp.github.io/CoreNLP/human-languages.html>

<https://opennlp.sourceforge.net/models-1.5>

<http://timm.ujaen.es/recursos/spanish-wordnet-3-0>

<http://spanishfn.org>

<https://catalog.ldc.upenn.edu/LDC2011T12>

http://clic.ub.edu/corpus/en/ancoraverb_es

⁷ [ELRA nebula y Catálogo de recursos de ELRA](#).

⁸ [Aprobación por el Consejo de Ministros](#) el 1 de marzo de 2022.

puede ser más o menos libre y abierto: desde la puesta a disposición como dominio público, hasta el requisito de solicitud previa y autorización o el establecimiento de restricciones y condiciones. Ello se articula, en muy buena medida, a través de las licencias que se elijan acordes al marco normativo vigente.

4 Los corpus lingüísticos de referencia en España

Cabe destacar los corpus textuales de carácter general desarrollados por la RAE, centrando ahora la atención en el español. Así, el [Corpus de Referencia del Español Actual](#) (CREA), el [Corpus Diacrónico del Español](#) (CORDE) y el [Corpus del Español del Siglo XXI](#) (CORPES XXI), liderados por la Real Academia Española (RAE); y el modelo MarIA. Desde 2001, la RAE dispone del Corpus Científico-Técnico (CCT), formado por un núcleo de 3 millones de palabras y complementado con otras publicaciones científicas (Ballester Carrillo, 2004). También dispone de algunos repertorios terminológicos extraídos de diccionarios y glosarios.⁹

Desde 2015, a partir del Plan de Impulso de Tecnologías del Lenguaje de la SETSI (2015b), se generaron dos corpus de gran valor para el desarrollo de la IA en español. Por un lado, el corpus [CAPITEL](#), de propiedad conjunta de la Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA) y la RAE, y realizado en colaboración, es el mayor corpus anotado en español existente en la actualidad, con 300 millones de palabras. Fue desarrollado a partir de subcorpus originales de noticias sin anotar proporcionados por diversos medios de comunicación. Por otro lado, se generó el corpus de la Biblioteca Nacional Española a partir de sus recursos del Archivo de la Web Española (colección formada por los sitios web que se recolectan con el fin de preservar el patrimonio documental español en internet y asegurar el acceso al mismo). Este corpus se emplea para la generación del [modelo de lenguaje MarIA](#), en colaboración con el Barcelona Supercomputing Center.

Y tras la ENIA y los planes de digitalización de 2021, el PERTE de 2022 da clara continuidad y profundidad a los corpus CREA y CORPES XXI y al modelo MarIA. Así, el eje transversal de la política de creación de bases de conocimiento tiene como uno de los dos proyectos tractores el de “Creación de nuevos corpus amplios para la nueva economía de la lengua” (Proyecto 1, MAETD, 2022a, p. 11). Se recuerda que, además de un buen uso de la IA, se requiere un buen corpus con gran cantidad de datos y bien diseñado, y también corpus específicos en otras lenguas también oficiales en España. Este proyecto tractor incluye el proyecto [Lengua Española e Inteligencia Artificial](#) (LEIA), impulsado por la RAE con la SEDIA, con las 22 academias de los distintos países hispanoamericanos y con fuerte colaboración público-privada. Las actuaciones incluyen, también, los medios de comunicación, así como las aportaciones al arriba referido corpus de la Biblioteca Nacional de España.

5 Políticas de tecnologías y recursos lingüísticos y el papel del sector público

La enorme potencialidad de las tecnologías del lenguaje, con la IA a la cabeza, y de los recursos lingüísticos ha ido ocupando un lugar relevante en los últimos años. Ya se han mencionado instrumentos desde el libro blanco de Melero et al. (2012), en especial el *Informe* y el *Plan* de la SETSI (2015a, 2015b) el *Estudio de caracterización* de la SESIAD (2018). El Plan de Impulso de las Tecnologías del Lenguaje, de 2015 y para 5 años, definió correctamente los objetivos, esencialmente (1) el desarrollo de infraestructuras lingüísticas en español y otras lenguas oficiales de España;¹⁰ (2) el impulso de la industria de las tecnologías

9 Como recurso más importante, el Corpus Científico-Técnico (CCT) se vale del Diccionario Esencial de las Ciencias de la Real Academia de Ciencias. También incluye el manual Merck de medicina, y se cuenta con el Diccionario Esencial de las Ciencias II de la Real Academia de Ciencias Exactas y Físicas y Naturales; el Diccionario médico etimológico de Montefiore, el glosario de Saludalia, la Neoloteca del Centre de terminologia TERMCAT, el *Diccionario terminológico y de siglas del subsector de lubricantes* de la Asociación Española de Lubricantes, el *Glosario médico multilingüe* de la DG III de la Unión Europea y el *Glosario oftalmológico* de Oftalmored. De igual modo, había convenio RAE-Consejo General del Poder Judicial (CGPJ) en el proyecto del *Diccionario de términos jurídicos*, así como participación de la Real Academia de Ingeniería, la Real Academia Nacional de Medicina, la Real Academia de Ciencias Exactas, Físicas y Naturales, la Real Academia Nacional de Farmacia y un largo etcétera de recursos (SETSI, 2015a, pp. 49, 57 y ss).

10 Así, el objetivo 1 de desarrollo de infraestructuras lingüísticas se concretaba en poner a disposición infraestructuras lingüísticas (recursos y procesadores) de propósito general; reducir la distancia que nos separa respecto del inglés; asegurar la disponibilidad pública gratuita o a bajo coste de infraestructuras lingüísticas de calidad, al menos a pymes; evitar duplicidades y buscar sinergias con los diversos agentes y niveles; disponer de herramientas comunes para la generación de recursos y campañas de evaluación; adoptar normas técnicas de interoperabilidad, una política de licencias adecuada y mecanismos de protección de datos personales en la generación de recursos lingüísticos; finalmente, potenciar métodos de generación automática de recursos lingüísticos.

del lenguaje, el fomento de la transferencia de conocimiento entre el sector investigador y la industria, así como la internacionalización.¹¹ Desde 2020, los planes e instrumentos no han dejado de producirse en el contexto digital, destacando, en julio de 2020, la Agenda Digital: España Digital 2025 (MAETD, 2020a, 2020b), que contiene un eje estratégico específico sobre la transformación digital del sector público; el Plan de Recuperación, Transformación y Resiliencia (MAETD, 2021a) y el Plan de Digitalización de las Administraciones Públicas 2021-2025 (MAETD, 2021b). A ello hay que añadir la Estrategia Nacional de Inteligencia artificial (ENIA) (MAETD, 2020c), en cuyo eje estratégico 3, medida 14, incluye impulsar las tecnologías lingüísticas y emplearlas en el sector público y situar la lengua española en una posición de liderazgo en el mundo de la IA.

A partir de ahí, cabe destacar el más concreto PERTE “Nueva Economía de la Lengua”, dotado con 1.100 millones de euros (30 exclusivamente para otras lenguas oficiales) y vinculado al Plan de Recuperación a través del eje transversal de la transformación digital. Sus medidas se han articulado en cinco ejes de carácter general (MAETD, 2022a, pp. 10-33). Como eje transversal, una base de conocimiento en español y otras lenguas oficiales; una IA en español (eje II), ciencia en español (eje III); aprendizaje del español y español en el mundo (eje IV) y el eje de industrias culturales. Por lo que ahora más interesa, el ya mencionado eje de base de conocimiento en español (96,7 millones) y, en especial, el eje II busca que “la inteligencia artificial procese adecuadamente en español, que ‘piense en español’” (MAETD, 2020a, p. 14) cuenta con 334,6 millones. En este eje II, el proyecto tractor 3, relativo al desarrollo de la IA en español y otras lenguas oficiales, señala la creación de modelos de lenguaje de alto valor generales y específicos (sanidad, justicia, etc.), hacer evolucionar el modelo MarIA y un test de referencia de evaluación de comprensión del lenguaje general similar –se afirma– a los GLUE/SUPERGLUE para el inglés. El proyecto tractor 4 supone la creación de la Red de Excelencia en IA, formación y cátedras. El proyecto tractor 5 se centra en el impulso de la industria de la IA en español, con financiación de proyectos relacionados con el lenguaje natural, desarrollo de herramientas, buenas prácticas, etc.

El papel del sector público en las políticas es central. Ya en el Plan de la SETSI (2015b, p. 19), el tercer objetivo era que la Administración pública fuera impulsora de la industria del lenguaje a través de la generación, estandarización y difusión de recursos lingüísticos creados en el contexto de la actividad de gestión pública propia de la Administración). El mismo fue ampliamente desarrollado a través de cuatro ejes de medidas.¹²

Se busca, en concreto, incorporar las tecnologías de la lengua al servicio público para mejorar la calidad y la capacidad del servicio público; lanzar y compartir proyectos innovadores; facilitar el desarrollo de nuevos componentes y el análisis de grandes corpus documentales; crear plataformas comunes de tecnologías de la lengua para compartir recursos eficientemente, con reutilización, código abierto, interoperabilidad y conocimiento compartido, y compartir servicios con otras instituciones del sector (RAE, Instituto de España, organismos iberoamericanos, etc.) En particular, respecto a los recursos lingüísticos, se tiene en cuenta el gran valor potencial que tiene buena parte de la información que genera el sector público como recurso lingüístico. En este ámbito, se subraya la necesaria coordinación para el desarrollo de recursos, herramientas comunes para su generación y evaluación, todo con normas técnicas de interoperabilidad, una política de licencias adecuada y mecanismos de protección de datos personales.¹³ En varios documentos se recuerda que, durante años, el elemento más descargado del portal de datos abiertos de la UE fue el corpus paralelo de traducciones del Parlamento Europeo.¹⁴

11 En cuanto al objetivo del impulso de la industria de las tecnologías del lenguaje, se busca mejorar la visibilidad y la transferencia de conocimiento del sector del procesamiento del lenguaje natural y de la traducción automática; mejorar el tejido investigador en la industria y mejorar la internacionalización de las empresas del sector, especialmente en el mercado iberoamericano y norteamericano; y apoyar a grandes empresas españolas asentadas en dichos mercados; estrechar la cooperación con la comunidad iberoamericana, y liderar allí la implantación de las tecnologías de la lengua.

12 El papel del sector público es esencial y las medidas se estructuran en los ejes siguientes: una buena gobernanza (eje 0); apoyo al desarrollo de infraestructuras lingüísticas (eje I); impulso de la industria del lenguaje (eje II); la Administración como impulsor de esta industria (eje III), y lanzamiento de “proyectos faro” (eje IV), que permitan la sinergia con sectores estratégicos, que tengan impacto, generen recursos reutilizables y adquirir experiencia y conocimiento.

13 De especial interés, SETSI, 2015a, pp. 127 y ss.; apartado 8 (“Las Administraciones Públicas y el PLN y TA en los servicios públicos”) y en particular el 8.3.

14 Actualmente puede descargarse desde [la web de la Comisión Europea](#).

6 La concurrencia y variabilidad de la regulación aplicable a los recursos lingüísticos

Las tecnologías y recursos lingüísticos quedan atraídos especialmente por la regulación del ámbito del derecho de propiedad intelectual y el ámbito afín del derecho de la reutilización de datos (Cerrillo i Martínez y Xalabarder, 2018), a los que se va a hacer especial referencia. La IA presenta enormes retos para la propiedad intelectual (Comisión Europea, 2022). De todos modos, no hay que excluir, en su caso, la aplicación del régimen de la propiedad industrial (SETSI, 2015a, p. 35) y META-NET.¹⁵ Y especialmente respecto de los datos y el procesamiento de los mismos con sistemas de IA, dada la dudosa regulación y protección aplicable a los mismos, hay que tener en cuenta la posible aplicación de sistemas de protección de secreto industrial.¹⁶ De igual modo para el sector público, que aquí es el que interesa especialmente, cabe tener en cuenta el derecho de la contratación pública. En modo alguno es sencilla la concurrencia de estos regímenes jurídicos.

Las políticas, tareas y fases del desarrollo de recursos lingüísticos quedan condicionadas por la regulación. En cualquier caso, el régimen jurídico es muy variable en razón del origen, fuente y naturaleza tanto de los recursos o datos y de las herramientas como de los actores implicados. Así, las fuentes de estos corpus lingüísticos o las bases de datos pueden ser muy diversas (titularidad pública, privadas, así como Administraciones públicas que custodian contenidos, como la Biblioteca Nacional, etc.) y de naturaleza muy variada (textos científicos, documentos traducidos y memorias de traducción, contratos, sentencias, historias clínicas, patentes, textos anotados, noticias, etc.), listas de nombres (de personas, organizaciones, marcas, topónimos, etc.), taxonomías y clasificaciones, o terminologías y diccionarios de titularidad pública o privada; asimismo, pueden provenir de fuentes distintas (Cerrillo i Martínez y Xalabarder, 2018, p. 19). También la procedencia de las herramientas y sistemas de procesamiento y tratamiento de estas fuentes pueden ser muy variados tanto por la titularidad como si es resultado de fórmulas de contratación o convenios, o bien de financiación más o menos pública.

7 Los titulares de los recursos lingüísticos tienen el derecho sui géneris a que no se pueda hacer minería de datos sin su autorización, salvo excepciones

Como punto de partida hay que tener en cuenta el régimen de la propiedad intelectual, en concreto, el – complejo– régimen de las bases de datos. Y es que los datos en sí mismos no están protegidos por propiedad intelectual, por lo que cabe acudir a la protección que tiene la estructuración de los datos. Y en el caso concreto de las palabras sueltas, estas no se consideran creación intelectual protegible: “las palabras clave, sintaxis, comandos y combinaciones de comandos, opciones, valores predeterminados e iteraciones consisten en palabras, figuras o conceptos matemáticos que, considerados aisladamente, no son, como tales, una creación intelectual del autor del programa de ordenador” (STJUE, de 2 de mayo de 2012, párr. 66).¹⁷

La normativa, y más con la Directiva 2019/790, ha dotado a las bases de datos de cierta protección, llamada *sui géneris*. Así, las bases de datos se protegen como algo que no son propiamente creaciones intelectuales, pero se les dota de cierta protección de propiedad intelectual por el hecho de que son creadas como consecuencia de una inversión sustancial. Así, las bases de datos, colecciones de datos y otros recursos en general constituyen creaciones intelectuales por su selección o disposición (art. 12 y libro II, título VIII, arts. 133-137 del Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual [LPI]). Esta protección *sui géneris* es, en general, aplicable a un recurso lingüístico, aunque la propia base de datos no sea una creación intelectual original (ELRA, 2019; SETSI, 2015a, p. 39). Asimismo, para aplicar esta protección, basta que sea una compilación o recopilación estructurada de información, y no hace falta que sea una base de datos según habitualmente se entiende en informática. En consecuencia, el titular o fabricante tiene el derecho *sui géneris* a autorizar o no el uso, la extracción, la explotación o la

15 Ahí se afirmaba, en 2012, que podrían patentarse algunas invenciones tecnológicas (máquinas de traducción automática, algoritmo de reconocimiento de habla, por ejemplo), incluso los procesos de producción del recurso, incluidos los procedimientos tecnológicos. Hay remisión a *Informe “META-NET: A network of Excellence forging the Multilingual Europe Technology Alliance*, del 24 de febrero de 2012, que no ha podido ser localizado.

16 En este sentido, hay que tener en cuenta la [Directiva \(UE\) 2016/943](#) del Parlamento Europeo y del Consejo, de 8 de junio de 2016, relativa a la protección de los conocimientos técnicos y la información empresarial no divulgados (secretos comerciales) y, en España, la [Ley 1/2019](#), de 20 de febrero, de Secretos Empresariales (Girona, 2018).

17 Sentencia del Tribunal de Justicia de las Comunidades Europeas (Gran Sala), de 2 de mayo de 2012, C-406/10, SAS Institute contra World Programming Ltd.

“minería de textos y datos” (Directiva (UE) 2019/ 790 art. 2.2) de partes sustanciales o la totalidad de sus contenidos. Los usuarios de una base de datos solo pueden hacer una “normal utilización” de “sólo una parte” de los mismos (art. 34 de la LPI). Así pues, para el minado de datos, el procesamiento y el tratamiento con tecnologías disruptivas de recursos lingüísticos, en principio, debe contarse con la autorización o licencia de los titulares y según sus condiciones. (Para el ámbito específico de recursos lingüísticos, Tsiavo, et al., 2014).

La Directiva 2019/790 afirma el principio de bases de datos *minables por defecto*. Ello es algo engañoso, por cuanto sí que se parte de que el titular de la base de datos tiene derechos, si bien expresamente tiene que reservar los mismos. Sobre esta base, hay que poner especial atención en las excepciones para la minería de textos y datos, que pueden permitir que legalmente se pueda minar datos protegidos por derechos de autor. Estas excepciones se dan en los artículos 3 y 4 de la Directiva 2019/790 (Margoni y Kretschmer, 2022). La tendencia normativa es ampliar las excepciones a este derecho. Para que no se necesite esta autorización o licencia del titular de la base de datos, hay que acudir a la general excepción de investigación científica (art. 3 de la Directiva 2019/790 y art. 34.2.b de la LPI). Pero lo cierto es que estas excepciones son bastante estrictas. Como punto de partida, la tendencia jurisprudencial no es especialmente expansiva a reconocer esta excepción (Vicente Domingo, 2016; Martínez Martínez, 2018). El artículo 3 de la Directiva 2019/790 intenta ampliar esta excepción a organizaciones de investigación y a las instituciones del patrimonio cultural que copian material al que se tenga acceso legal para efectuar minería de datos con fines de investigación científica y almacenar copias (arts. 7 y 4). En estos casos, los titulares de las bases de datos no pueden imponer prohibiciones expresas, como sí pueden en el resto de los casos (art. 4).

Desde la doctrina se ha señalado que “las ‘sombras’ en esta regulación son bastante más ‘alargadas’ que sus ‘luces’” (Jiménez Serranía, 2020, p. 256), pues su alcance real es muy limitado. Así, esta excepción solo alcanzaría a la actividad automatizada, pero no al preprocesamiento o etiquetado de datos que sea realizado por humanos, lo cual es muy habitual. Además, quedarían fuera de la excepción las empresas del sector que no son organismos e instituciones del patrimonio cultural, que, para gozar de esta excepción, deben negociarla e incluirla en la contratación (Aurelius, 2022). De este modo, en el Derecho de la UE no se goza de las ventajas y posibilidades que se dan en otras regiones del mundo, como las relativas al *fair use* de Estados Unidos que permite un desarrollo mucho más fácil en el ámbito de la minería de datos, la IA y las tecnologías y recursos lingüísticos o la reciente excepción flexible que también se ha introducido en Japón (Leistner y Antoine, 2022). Se afirma por ello que “si la Unión Europea continúa cinchando de esta manera tan rígida al potro salvaje de la minería de datos probablemente el jinete empresarial acabe por tierra y maltrecho” (Jiménez Serranía, 2020, p. 256).

Si, por un lado, estas excepciones son insuficientes hoy por hoy pese a afirmarse la minería por defecto, por otro lado, la aplicación efectiva del derecho genera asimetrías que pueden perjudicar especialmente a las acciones públicas en materia de recursos lingüísticos. Y es que no puede desconocerse que, en la realidad, es muy difícil, para el titular de las bases de datos y recursos, probar que se ha incumplido una prohibición de minería de datos y que este incumplimiento le ha generado daños. En consecuencia, puede pensarse que iniciativas desde el sector privado quizá estén realizando minería de datos y textos que no es legal, pero que tiene muy bajo riesgo de que sea detectada de algún modo. Sin embargo, en el caso del sector público, las iniciativas y actividades que se emprenden en materia de recursos lingüísticos, necesariamente han de respetar el ordenamiento jurídico. En consecuencia, podría concluirse que la legalidad solo se aplica como una barrera para el sector público.

Y lo que es aún peor, para el caso español se ha regulado mal este débil régimen privilegiado de la excepción científica aplicada a la minería de datos. Así, la transposición de la Directiva 2019/790 al derecho español ha obviado la especialidad para organismos de investigación y responsables del patrimonio cultural (art. 67 del Real Decreto-Ley 24/2021, de 2 de noviembre, de transposición de directivas de la Unión Europea en determinados ámbitos, Aurelius, 2020). La técnica legislativa de este amplio artículo es deficiente, pero lo cierto es que el apartado 2.º permite, también en el caso de organismos de investigación o responsables de patrimonio, que los titulares de derechos de las bases de datos (en nuestro caso, los corpus) hayan reservado expresamente la posibilidad de hacer tratamientos de minería de textos y datos. Así pues, la actual y defectuosa transposición española parece ir en contra de la propia directiva, así como de las políticas y la [Estrategia](#)

[Europea de Datos](#) y la [Recomendación sobre un espacio común europeo de los datos para el patrimonio cultural](#), de 10 de noviembre de 2021.

Algún paso más se da con el [Reglamento de datos](#)¹⁸ que se aprueba en 2022. La nueva norma facilita el acceso a bases de datos que contienen información procedente de dispositivos y objetos de la internet de las cosas, además de la utilización y minería de dichas bases de datos. Así, se excluye que los titulares de estos datos puedan reclamar sus derechos sui generis respecto de bases de datos (considerando 84, art. 35). También se incluyen algunas medidas quizá aplicables en el contexto lingüístico para que los usuarios de dispositivos conectados puedan acceder a los datos generados por ellos. Igualmente, se incluyen incentivos para que los fabricantes sigan invirtiendo en la generación de datos de alta calidad.

Cabe mencionar otras vías que en algunos supuestos pueden flexibilizar los derechos de los titulares de bases de datos y recursos lingüísticos y facilitar, en su caso, su explotación. Así, la Directiva 2019/790 también prevé el uso de obras y otras prestaciones fuera del circuito comercial por parte de las instituciones responsables del patrimonio cultural (arts. 8 y ss.), así como el mecanismo de las licencias colectivas ampliadas en razón del artículo 12, que pueden conceder las organizaciones de gestión colectiva, a partir del modelo escandinavo. Quizá en algún caso, en el contexto de recursos y tecnologías lingüísticas, cabría aplicar la posibilidad de explotar bases de datos que pongan a disposición “obras fuera del circuito comercial”, en el caso de instituciones responsables del patrimonio cultural (art. 71.7 y 9 del Real Decreto-Ley 24/2021). También en algún caso podría aplicarse la posibilidad de explotar bases de datos para fines de conservación del patrimonio cultural conforme al artículo 37 de la LPI (art. 69.3 del Real Decreto-Ley 24/2021). En todo caso, se trataría de supuestos específicos.

Pero, por otro lado, la Directiva 2019/790, en los artículos 15 y 16, reconoce un nuevo derecho exclusivo a favor de los editores de prensa, quienes pueden impedir el uso en línea de sus publicaciones durante los dos años siguientes a la publicación. Ello, en su caso, puede impedir algunas prácticas de minería de estos datos y repertorios lingüísticos.

8 Las exigencias de la regulación de reutilización y la interoperabilidad de los recursos lingüísticos

No es lugar ahora para sintetizar la esencial y dinámica acción de la UE y de España respecto de la reutilización de datos y los datos abiertos (Moro et al., 2020; Valero Torrijos, y Martínez Gutiérrez, 2022). La [Directiva \(UE\) 2019/1024](#), del Parlamento Europeo y del Consejo, de 20 de junio de 2019, relativa a los datos abiertos y la reutilización de la información del sector público, está esencialmente transpuesta en España por la Ley 37/2007, de 16 de noviembre, sobre la Reutilización de la Información del Sector Público y, más recientemente, por el Real Decreto-Ley 24/2021. En general, la legislación europea y la española implican el fomento de la difusión y reutilización de datos y, por lo que ahora interesa, de los recursos lingüísticos generados por el sector público. Ahora bien, no hay una obligación estricta de puesta a disposición que permita la reutilización para fines comerciales o no comerciales, en nuestro caso, la libre explotación con tecnologías lingüísticas (Beltrán, 2022). En cada regulación concreta o en cada caso concreto, las Administraciones titulares promotoras de los recursos lingüísticos deben determinar los modelos para su explotación y acceso, desde la directa puesta en dominio público, el establecimiento de restricciones en las licencias o la necesidad de solicitud previa y la consiguiente decisión. La normativa no obliga a la apertura de estos recursos, pero si el sector público decide abrir estos recursos para su reutilización, las condiciones que en su caso establezcan han de ser claras, justas y transparentes, no discriminatorias, no deben restringir las posibilidades de reutilización ni limitar la competencia (art. 4 de la Ley 37/2007).

Para el desarrollo de tecnologías y recursos lingüísticos, la interoperabilidad es también un elemento esencial (SETSI, 2015a, p. 41). De especial relevancia es la estandarización y normalización técnica de la codificación, la anotación y el etiquetado de los corpus, especialmente de los metadatos. Es así como puede hacerse efectiva la (re)utilización de datos, la combinación de herramientas y recursos lingüísticos. Para

¹⁸ Propuesta de Reglamento sobre normas armonizadas para un acceso justo a los datos y su utilización (Ley de Datos), de 23 de febrero de 2022.

facilitar la interoperabilidad, se han propuesto esquemas que engloban múltiples niveles de anotación.¹⁹ El organismo ISO ha reconocido varias de las propuestas, sin embargo, dada la dinamicidad e inestabilidad del sector, no existe una única anotación lingüística. El Real Decreto 4/2010, de 8 de enero, por el que se regula el Esquema Nacional de Interoperabilidad (ENI) en el ámbito de la Administración Electrónica establece los elementos mínimos a seguir, si bien no parece alcanzar el grado de detalle concreto para los recursos lingüísticos.

9 Hacia las licencias permisivas de la explotación de recursos lingüísticos: las licencias permisivas del ecosistema Meta-Share y su proyección a los corpus españoles

Según se ha analizado, el punto de partida es que el titular del recurso lingüístico puede decidir su explotación y poner condiciones, con los límites normativos. Son muchos los impulsos que conducen a que el sector público tienda cada vez más hacia la apertura y reutilización de sus datos y recursos, máxime cuando los genera precisamente como resultado de sus políticas digitales y de IA (Cotino, 2020). La mejor solución pasa por elegir correctas licencias de uso y explotación de tecnologías y repertorios lingüísticos.

En 2015, desde el Gobierno (SETSI, 2015a, pp. 37-38) se afirmaba que las licencias “han de incluir el derecho a consultar el contenido de cada recurso lingüístico, pero también los de transformación, difusión y distribución de dichos recursos o de sus derivados, siempre de acuerdo con las necesidades y las políticas de los/las creadores/as y usuarios/as de los recursos”. Pese al principio de “minable por defecto”, resulta de especial interés que el titular de un recurso lingüístico exprese la posibilidad de hacer minería de datos y tratamientos automatizados con las bases de datos, corpus y otros recursos lingüísticos.

Cada desarrollador puede *crear* su propia licencia. Y la tendencia hacia la apertura se aprecia en el caso de las licencias de *software*. Así, en 2015 ya había más de 2.000 tipos de licencias de *software* diferentes. En todo caso, el 90 % de los proyectos de *software* abierto usan unos diez tipos de licencias (SETSI, 2015a, p. 38), que básicamente se reconducen a dos tipos de licencias: las “permisivas” (un 42 % de los casos), que establecen pocas condiciones (licencias MIT, BSD y Apache License 2.0) y las conocidas licencias permisivas de “*copyleft*” (GPL y LGPL, un 37 %, frente al *copyright*) (SETSI, 2015, p. 38), por las que, en general, el *software* derivado de otro *software* debe ser distribuido con la misma licencia que el original (O’Grady, 2014).

Hay una tendencia hacia la apertura, tanto de recursos como de *software* y tecnologías lingüísticas, sobre la que, quizá exageradamente, se afirma que es “exponencial” (SETSI, 2015a, p. 46). Cabe apostar por licencias permisivas. Hay muchas tecnologías de procesamiento lingüístico abiertas²⁰ (tokenización, segmentación morfológica, segmentador de frases, análisis sintáctico, identificación y clasificación de entidades nombradas, la resolución de la correferencia, etc.) que incluyen algoritmos de aprendizaje automático.

Como se ha adelantado, en el contexto de ELRA, la iniciativa europea [Meta-Share](#) ha estado, desde 2013, en la base de las políticas digitales lingüísticas y ha creado un catálogo de recursos lingüísticos en Europa con la filosofía de la apertura, acceso y reutilización. Ahí puede accederse a un [conjunto de licencias](#) inspiradas en las licencias ELRA que incorporan las nuevas tendencias permisivas y recursos libres en la órbita de Creative Commons. Así, se proporcionan recursos para infraestructura abierta, distribuida, segura e interoperable en el ámbito de las tecnologías lingüísticas. Ello es de especial utilidad para la infraestructura de proveedores y los usuarios de tecnologías y recursos lingüísticos, así como para los integradores y vendedores de estos recursos. Así, se proponen ocho diferentes conjuntos de plantillas de licencias. Todas ellas orientadas hacia la apertura y el ecosistema propicio para las tecnologías del lenguaje.

Las mismas se organizan en diversos niveles o ejes. El punto de partida es que el titular se reserva todos los derechos, y estas licencias implican diversos niveles de renuncia a los mismos. Así, se pueden dar recursos abiertos en los que no hay reserva, esto es, sin ninguna condición ni restricción; también licencias con las

¹⁹ Los más conocidos son ATLAS, PAULA, GATE, UIMA, NIF (con el objetivo de utilizar principios y paradigmas *open link data* para anotar información lingüística), también formatos KAF y NAF.

²⁰ [Natural Language Toolkit \(NLTK\)](#), [SpaCy](#), [TextBlob](#), [Textacy](#), [PyTorch-NLP](#), [Retext](#), [Compromise](#), [Natural](#), [Nlp.js](#), [OpenNLP](#), [Stanford CoreNLP](#), [CogCompNLP](#). Pueden seguirse estas tecnologías y herramientas en Barker (2019). En español, se hablaba en 2015 de [Freeling](#), si bien es difícil señalar su relevancia actual. También SETSI, 2015a, pp. 40-41.

diversas variantes de Creative Commons. Otro eje de licencias [Meta-Share Commons](#) permite que los recursos estén disponibles solo para otros miembros de la Alianza Meta para el uso y la explotación de los recursos, pero bajo el control de los propietarios sobre su distribución. Finalmente, un tercer eje de licencias son algo más restringidas y tienen diversas condiciones, si bien se permite que recursos cerrados también puedan llegar a esta comunidad. De igual modo, esta red Meta ofrece plantillas de documentos legales de extraordinaria utilidad (*Acuerdos del creador del recurso*, *Memorando de entendimiento para los miembros de la red*; *Acuerdo de nivel de servicio*). Estos documentos definen las condiciones de uso para fines de investigación y comerciales.

También hay que tener en cuenta las herramientas desarrolladas por el [Linguistic Data Consortium](#) (s.f.) (AGTK, CTK, LDC Word Aligner, herramientas de conversión SPHERE, XTrans) y recursos desde de la LREC (en el contexto de ELRA-ELDA y CLARIN).

Así pues, desde hace una década se cuenta con la experiencia de compartición de recursos y de licencias, si bien es cierto que su complejidad requiere un mayor conocimiento desde todos los actores, especialmente las pymes, parejo al desarrollo normativo que también se da. También es cierto que, en ocasiones, los modelos de licencias abiertos pueden dificultar la inversión; ello sucede en la medida en la que el usuario de estos recursos o tecnologías luego también tiene que sujetarse a este tipo de licencias abiertas respecto de los desarrollos y datos generados, lo cual puede limitar la explotación de sus productos.

10 Para concluir: la necesidad de planes y modelos de explotación y sostenibilidad de los recursos lingüísticos en España

A lo largo del estudio se ha expuesto que las ingentes cantidades de datos contenidos en los textos de webs, plataformas, bases de datos, medios, etc., son la materia prima, el *big data* que es procesado a través de la inteligencia artificial, la minería de datos y otras tecnologías disruptivas lingüísticas para el procesamiento del lenguaje natural y la traducción automática. Es indudable el potencial cultural, social, económico y estratégico que tienen para el español y otras lenguas españolas. Desde hace una década, los poderes públicos en España han fijado los objetivos de las políticas en IA y lenguas; y también en este estudio se expone el especial papel del sector público. Y en los últimos años se han intensificado mucho estas políticas (ENIA, PERTE “Nueva Economía de la Lengua” de 2022). Según se ha insistido, resulta clave la generación (pública) de infraestructuras y recursos lingüísticos, bases de conocimiento y, especialmente, corpus lingüísticos que *alimentan* a la IA y otras tecnologías lingüísticas desarrolladas especialmente por el sector privado. El estudio ha abordado la regulación aplicable a estos recursos lingüísticos, como elemento que condiciona las políticas, el planeamiento y los modelos de explotación y sostenibilidad de los mismos. Especialmente se centra la atención en el enfoque de la propiedad intelectual y el régimen de reutilización de datos. Se ha expuesto cómo los promotores y titulares de los corpus y otros recursos lingüísticos no están claramente obligados a abrir estos recursos y tienen el derecho sui generis a que no se pueda hacer minería de datos u otros procesamientos sin su autorización. Se ha examinado la normativa europea y española más reciente, con las excepciones a favor de la investigación, en su caso aplicables al ámbito de la lengua, otras posibles excepciones y la tendencia normativa hacia la flexibilidad y apertura a la que tiende el Derecho de la UE. En cualquier caso, la articulación de estas políticas de apertura a la explotación de recursos lingüísticos pasa por la elección de licencias permisivas, que son la clara tendencia. Así, especialmente se han de tener en cuenta las licencias Meta-Share así como las herramientas y recursos desde el ecosistema de ELRA, y aprovechar en lo posible su uso respecto de los corpus y otros recursos lingüísticos que se promueven en España.

Para concluir, cabe exponer los elementos clave en la planificación y adopción de modelos de explotación y sostenibilidad de los recursos lingüísticos en España. Así, a la vista del contexto jurídico, la generación de recursos lingüísticos como los que se desarrollan en España requieren de una estrategia para la puesta a disposición a partir del marco normativo. Esta puesta a disposición se articula esencialmente en la elección de licencias para la puesta a disposición de investigadores, industria y ciudadanía.

Un adecuado plan y modelo de explotación y sostenibilidad de los recursos lingüísticos es cuestión condicionada por el derecho, pero que va más allá del mismo y alcanza disciplinas como las políticas públicas, la economía, así como las tecnologías de la información y de sistemas. Se precisa la integración de recursos lingüísticos en catálogos, la generación y la resolución de problemas de propiedad intelectual antes de su

distribución. A partir de ahí se pasa a la distribución y difusión de los recursos con sus diferentes usos. Con un plan o modelo de gestión se debe cubrir todo el ciclo de vida de los datos, e implica la descripción de la política a llevar a cabo y la sostenibilidad de los datos y su explotación. Deben tenerse en cuenta las diferentes dimensiones²¹ y factores y medios de impacto identificados; los mismos incluyen la producción y el uso de los recursos lingüísticos, la concesión de licencias, el mantenimiento y la conservación, las infraestructuras para los recursos lingüísticos, la identificación y el intercambio de recursos, la evaluación y validación, la interoperabilidad y las cuestiones normativas.

El desarrollo del plan y del modelo implica diversas tareas: analizar bien los datos del corpus o recurso de que se trate; determinar los actores en el ecosistema y en la cadena de valor de la explotación de datos (desde particulares hasta instituciones educativas y de investigación, responsables políticos, agencias de financiación, pymes y grandes empresas, proveedores de servicios y medios). En España se cuenta con unos 200 agentes del sector de tecnologías del lenguaje, asociaciones y grupos de investigación.²²

Es necesario analizar y tener en cuenta los recursos necesarios a invertir, los recursos disponibles y la estimación de resultados y retornos con la explotación de datos con relación a los sujetos de la cadena de valor, esto es, el valor potencial generado por las actividades de explotación de datos. Hay que elegir un modelo de distribución gratuita o de pago dependiendo del tipo de usuario/entidad (gran empresa, pymes, investigador, etc.) y/o del fin para el que se vaya a utilizar el recurso (investigación, uso comercial, etc.). Se han de tener en cuenta las distintas posibilidades para la apertura y distribución de los corpus.

Sobre estas bases hay que formular un borrador para someterlo a aportaciones y comentarios tanto de los titulares o promotores de la base o corpus como de los agentes involucrados que deban hacer su explotación. Una vez adoptados el plan y el modelo de explotación, hay que implementarlo y hacer una evaluación y seguimiento del mismo para su corrección o reformulación.

Para todas estas acciones se cuenta con excelentes recursos de apoyo, como el modelo general de explotación de datos de Colombia (Gamez Daza, 2021). Y para el caso de recursos lingüísticos, cabe especialmente tener en cuenta recursos sobre explotación de datos de ELRA ([Data Management Plan](#)) y el modelo que sigue a resultados del proyecto FlaReNet (Calzonolari et al., 2012). Todas las fases están documentadas: descripción del proyecto, adquisición de datos, documentación de datos y metadatos, licencia de datos, conservación e intercambio de datos.

Estas acciones deben abordarse en el caso de CAPITEL (como se vio, el mayor corpus anotado de la RAE y la SEDIA), donde especialmente debe garantizarse a los medios de comunicación que el uso de los corpus sea exclusivamente como recurso lingüístico y que no se permita la republicación de las noticias que los componen. También deben acometerse estas acciones respecto del corpus para generar el modelo de lenguaje MarIA (Biblioteca Nacional Española, SEDIA, Barcelona Supercomputing Center) cuya apertura se pretende.

21 (a) documentación b) interoperabilidad c) disponibilidad, intercambio y distribución d) cobertura, calidad y adecuación e) sostenibilidad f) reconocimiento g) desarrollo h) infraestructura e i) cooperación internacional).

22 Un análisis exhaustivo del ecosistema en SESIAD (2018, pp. 28-59), y del modelo de negocio y mercado también en SESIAD (2018, pp. 60 y ss). Se identificaron 127 empresas y 63 centros de investigación. Cabe también recordar la existencia de la Sociedad Española para el Procesamiento del Lenguaje Natural ([SEPLN](#)), con casi 40 años, asociaciones como AETER, desde 1997, así como [34 grupos de investigación](#) españoles.

Referencias

- Alcalde Bezhold, Guillermo, y Alfonso Farnós, Iciar. (2019). Utilización de tecnología Big Data en investigación clínica. *Revista de Derecho y Genoma Humano. Genética, Biotecnología y Medicina Avanzada, Extra*, 55-83.
- Aurelius. (2021, 14 de noviembre). [La criticable incorporación de la excepción de minería de textos y datos al ordenamiento español en el Real Decreto Ley 24/2021](#). Lvcentinvs.
- Ballester Carrillo, Almudena. (2004). [El Corpus Científico-Técnico de la Real Academia Española](#). En Luis González y Pollux Hernández (coords.), *Las palabras del traductor. Actas del II Congreso “El español, lengua de traducción”* (pp. 129-136). Esletra.
- Barker, Dan. (2019). [12 open source tools for natural language processing](#). Opensource.
- Beltrán Castellanos, José Miguel. (2022). Alcance de la obligación, seguridad jurídica y responsabilidad en la reutilización de la información del sector público. En Julián Valero Torrijos y Rubén Martínez Gutiérrez (dirs.) (2022), *Datos abiertos y reutilización de la información del sector público* (pp. 127-157). Comares.
- Calzolari, Nicoletta, Bel, Nuria, Choukri, Khalid, Mariani, Joseph, Monachini, Monica, Odiijk, Jan, Piperidis, Stelios, Quochi, Valeria, oria, Claudia. (2012). [Final FLReNet Deliverable Language Resources for the Future – The Future of Language Resources](#). European Language Resources Association.
- Cerrillo i Martínez, Agustí, y Xalabarder, Raquel. (2018). El impacto del derecho en el uso de las tecnologías del lenguaje en las administraciones públicas. *Revista de Llengua i Dret, Journal of Language and Law*, 70, 17-30. <http://dx.doi.org/10.2436/rld.i70.2018.3159>
- Comisión Europea. (2022). [Study on copyright and new technologies: copyright data management and artificial intelligence](#). Oficina de Publicaciones de la Unión Europea. <https://data.europa.eu/doi/10.2759/570559>
- Cotino Hueso, Lorenzo. (2020a). Big data. En Benigno Pendás (ed.), [Enciclopedia de las Ciencias Morales y Políticas para el siglo XXI](#) (pp. 96-100). Real Academia de Ciencias Morales y Políticas-Boletín Oficial del Estado.
- Cotino Hueso, Lorenzo. (2020b). Ética, valores y principios del “open data” y los retos futuros de la apertura de datos públicos. *El Consultor de los Ayuntamientos y los Juzgados, Extra 3*, 147-166.
- Cotino Hueso, Lorenzo, y Todolí Signes, Adrián (coords.). (2022). *Explotación y regulación del uso del big data e inteligencia artificial para los servicios públicos y la ciudad inteligente*. Tirant lo Blanch.
- European Language Resources Association. (2019, 24 abril). [What’s new in the Directive on Copyright in the Digital Single Market](#). *ELRC+3 Newsletter*.
- Gamez Daza, Luis Segundo (coord.). (2021). [Modelo de explotación de datos para las entidades públicas](#). Departamento Nacional de Planeación.
- Girona Domingo, Ramón Miguel (2018). Hacia una nueva protección de los secretos industriales y comerciales. La Directiva 2016/943 del Parlamento Europeo y del Consejo. *Revista Jurídica de la Comunidad Valenciana*, 66, 5-20.
- Jiménez Serranía, Vanessa (2020). Datos, minería e innovación: qvo vadis, Europa? Análisis sobre las nuevas excepciones para la minería de textos y datos. *Cuadernos de Derecho Transnacional*, 12(1), 247-258, <https://doi.org/10.20318/cdt.2020.5188>
- Leistner, Matthias, y Antoine, Lucie. (2022). [IPR and the use of open data and data sharing initiatives by public and private actors](#). Parlamento Europeo.
- Linguistic Data Consortium. (s.f.). [Tools](#).

- Ministerio de Asuntos Económicos y Transformación Digital. (2020a). [*Agenda Digital: España Digital 2025*](#).
- Ministerio de Asuntos Económicos y Transformación Digital. (2020b). [*Agenda Digital: España Digital 2025. Resumen ejecutivo*](#)
- Ministerio de Asuntos Económicos y Transformación Digital. (2020c). [*Estrategia Nacional de Inteligencia Artificial*](#).
- Ministerio de Asuntos Económicos y Transformación Digital. (2021a). [*Plan de Recuperación, Transformación y Resiliencia*](#).
- Ministerio de Asuntos Económicos y Transformación Digital. (2021b). [*Plan de Digitalización de las Administraciones Públicas 2021-2025*](#).
- Ministerio de Asuntos Económicos y Transformación Digital. (2022a). [*PERTE Nueva Economía de la Lengua. Memoria Técnica*](#).
- Ministerio de Asuntos Económicos y Transformación Digital. (2022b). [*PERTE Nueva Economía de la Lengua. Resumen Ejecutivo*](#).
- Melero, Maite, Badía, Toni, y Moreno, Asunción. (2012). [*La lengua española en la era digital*](#). Springer.
- Margoni, Thomas, y Kretschmer, Martin, (2022). A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. *GRUR International*, 71(8), 685-701. <https://doi.org/10.1093/grurint/ikac054>
- Martínez Martínez, Nuria. (2018). [*El límite de la ilustración con fines educativos y de investigación científica a la propiedad intelectual*](#) [Tesis doctoral]. Universidad de Alicante.
- Ministerio de Asuntos Exteriores. (2022). [*El español en el mundo*](#).
- Moro, M.^a Ascensión, Colón, Borja, y Magro, Roberto. (coords.) (2020). *El Consultor de los Ayuntamientos y los Juzgados, Extra 3*.
- O'Grady, Stephen. (2014, 14 de noviembre). [*What are the Most Popular Open Source Licenses Today?*](#). Redmonk.
- Ortega Giménez, Alfonso. (2019). Implicaciones jurídicas de la internalización de la tecnología del Big Data y Derecho Internacional Privado. *Revista de Derecho y Genoma Humano. Genética, Biotecnología y Medicina Avanzada, Extra*, 169-204.
- Rigau i Claramunt, German. (2022). La tecnología del lenguaje: la inteligencia artificial centrada en el lenguaje. En Instituto Cervantes, [*El español en el mundo 2022. Anuario del Instituto Cervantes*](#) (pp. 201-218).
- Secretaría de Estado de Telecomunicaciones y Sociedad de la Información. (2015a). [*Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España*](#). Ministerio de Industria, Energía y Turismo.
- Secretaría de Estado de Telecomunicaciones y Sociedad de la Información. (2015b). [*Plan de Impulso de las Tecnologías del Lenguaje*](#). Ministerio de Industria, Energía y Turismo.
- Secretaría de Estado para la Sociedad de la Información y la Agenda Digital. (2018). [*Estudio de caracterización del sector de tecnologías del lenguaje*](#). Observatorio Nacional de Telecomunicaciones y de la SI.
- Tasa Fuster, Vicenta. (2022). Oficialidad lingüística e inteligencia artificial: Una reflexión sobre las obligaciones lingüísticas de las administraciones públicas ante la inteligencia artificial. En Cotino Hueso, Lorenzo y Todolí Signes, Adrián (coords.), [*Explotación y regulación del uso del big data e inteligencia artificial para los servicios públicos y la ciudad inteligente*](#) (pp. 289-316). Tirant lo Blanch.

- Tsiavos, Prodromos, Piperidis, Stelios, Gavrilidou, Maria, Labropoulou, Penny, y Patrikakos, Tasos. (2014). [Legal Aspects of Text and Data Mining](#) *Legal Aspects of Text and Data Mining*, Proyecto QTLaunchPad, Wikibooks.
- Valero Torrijos, Julián, y Martínez Gutiérrez, Rubén. (dirs.) (2022). *Datos abiertos y reutilización de la información del sector público*. Comares.
- Vicente Domingo, Elena. (2016). [Los límites del derecho de cita e ilustración con fines educativos o de investigación científica](#). En Raquel de Román Pérez (coord.), *La propiedad intelectual en las universidades públicas: titularidad, gestión y transferencia* (pp. 113-141). Comares.
- Wikipedia. (2022). [Anexo: Idiomas por el total de hablantes](#).