MESA REDONDA

Tecnologías del lenguaje, Plan de impulso y Compra Pública de Innovación.

En los años noventa arrancó la digitalización masiva y hoy buena parte de la materia prima informativa se produce ya directamente en formato electrónico (textos, audio, vídeo, mediciones, etcétera).

D. DAVID PÉREZ FERNÁNDEZ
Gabinete del Secretario de Estado para la Sociedad de la Información

D. JUAN DE DIOS LLORENS GONZÁLEZ
Agenda Digital
Ministerio de Energía, Turismo y Agenda Digital

No obstante, esto se refiere esencialmente a una forma de almacenamiento de la información que si bien facilita operaciones muy importantes como su replicación o su transmisión; no implica que sea información que un ordenador pueda "comprender", asignar significado a sus elementos y gestionar. Por eso, aunque se trate de información en formato electrónico, deberemos distinguir entre información estructurada, que es la que está preparada para ser "comprendida" por sistemas informáticos (bases de datos), e información no estructurada, que es la destinada a ser comprendida por humanos (libros, fotografías, películas, música).

Sin embargo, el volumen de información no estructurada crece tan vertiginosamente que su aprovechamiento está ya fuera del alcance de las capacidades humanas. Hemos pasado de la desnutrición a la obesidad informativa. No somos capaces de sacarle partido a las ingentes cantidades de información que almacenamos o a las que tenemos acceso. Y buena parte de esta información es de contenido lingüístico (oral y escrito). Por ello, es acuciante la necesidad de explotar automáticamente este volumen de información que crece vertiginosamente.

En este artículo denominamos tecnologías del lenguaje a un conjunto diverso de tecnologías que van jalonando el camino hacia una compresión automática cada vez más profunda del lenguaje humano, es decir, las que permiten explotar automáticamente ("estructurar") la información no estructurada expresada en lenguaje humano.

Caben varias clasificaciones de estas tecnologías, pero a grandes rasgos podemos hablar de tecnologías de procesamiento de

lenguaje natural (PLN), de traducción automática o asistida por ordenador y de sistemas conversacionales.

El procesamiento lingüístico

El procesamiento se puede realizar en cualquiera de los niveles del estudio del lenguaje (fonológico, morfológico, sintáctico, semántico y pragmático), y las tareas de procesamiento pueden ser muy diversas, por ejemplo:

- Segmentación (*Tokenization*) de los elementos que aparecen en el texto (palabras, números, símbolos, frases, párrafos,...).

- Análisis morfosintáctico (POST o *Part-of-Speech Tagging*), en el que a cada elemento del texto se le asigna una categoría sintáctica (sustantivo, adjetivo, verbo,...), se le asigna su lema y se le añade información morfosintáctica (número, género, tiempo verbal,...). También aquí hay tareas como reconocer conjuntos de elementos que forman una unidad léxica ("hombre rana", "caer en la cuenta").

- Análisis sintáctico *(Syntactic parsing, Chunking, Dependency Analysis o Constituent Analysis)*, que consiste en la construcción de los árboles sintácticos, y se puede realizar conforme a diversas teorías y formulaciones.

- Identificación de la acepción empleada *(Word Sense Disambiguation)* a cada elemento que pueda tener varias acepciones.

- Identificación, clasificación y desambiguación de entidades (nombres propios – personas, lugares, organizaciones... – y expresiones numéricas – fechas, cantidades de dinero...–) nombradas en el texto (NERC o *Named Entity Recognition and Classification*).

- Asignación de funciones semánticas (SLR o *Semantic Role Labeling*), que consiste en identificar y clasificar los argumentos semánticos asociados con los verbos o elementos predicativos en el texto. Este nivel es mayor que el de los árboles sintácticos. Por ejemplo, es invariante respecto a la voz pasiva o activa.

- Correferencia *(Coreference)*, que es la identificación de cada una de las diferentes menciones al mismo objeto en el texto ("Marta aprobó el examen. La chica se lo merecía").

- Detección e interpretación de expresiones temporales, teniendo en cuenta que pueden ser relativas ("hace un mes").

- Detección de estructuras discursivas (discurso, escritura, conversación, evento comunicativo).

- Análisis de sentimientos (*Sentiment Analysis* u *Opinion Mining*), que trata de identificar la emoción asociada al texto (por ejemplo, aprobación o repulsa).

- Clasificación, caracterización y comparación automática de textos.

- Traducción automática *(Machine Translation)* o asistida por ordenador.

Hay que recordar que se trata de realizar estas tareas automáticamente.

A pesar de la dificultad de este objetivo, el grado de desarrollo alcanzado por las tecnologías del lenguaje permite ya multitud de aplicaciones de utilidad, y su rápido desarrollo en los últimos años augura resultados cada vez más sorprendentes. Herramientas tales como buscadores de Internet, asistentes personales en los móviles, predictores y correctores automáticos de texto o traductores automáticos se han convertido en habituales para el desarrollo de nuestra labor cotidiana, sea cual sea el campo de actividad.

Pero hay otras muchas aplicaciones donde las tecnologías del lenguaje pueden resultar críticas para dotar al ciudadano de nuevos servicios avanzados, para ayudar a las organizaciones, y entre ellas a las administraciones, a optimizar muchos de sus procesos y obtener conocimiento muy valioso de su propia información y de aquella disponible en un mundo cada vez más digitalizado e interconectado. Cualquier paso en la mejora de la explotación automática de la información no estructurada de contenido lingüístico genera valor para la sociedad y resulta de aplicación transversal a todos los sectores productivos.

En consecuencia, las tecnologías del lenguaje tienen el potencial de generar un sector industrial emergente, innovador y transversal.

Un ejemplo de uso: Vigilancia sectorial

Para ilustrar la utilidad del procesamiento de lenguaje natural podemos mostrar los resultados ya obtenidos en el proyecto de Vigilancia Sectorial de la Secretaría de Estado para la Sociedad de la Información y Adenda Digital (SESIAD). Este proyecto responde a dos necesidades de la SESIAD, que son obtener un conocimiento profundo y actualizado del sector de las tecnologías de la información y las telecomunicaciones (TIC) y de la Sociedad de la Información, y su prospectiva, para la dirección de políticas públicas, y auxiliar a los evaluadores de solicitudes de ayudas públicas a cotejarlas con el corpus de ayudas concedidas y con las demás solicitudes presentadas.

En el contexto de este proyecto se han empleado diversas técnicas de procesamiento de lenguaje natural. Centrémonos en una de ellas, el modelo generativo estadístico de semántica latente denominado "Latent Dirichlet Allocation" (LDA).

A grandes rasgos, esta técnica permite caracterizar una colección de documentos (corpus) en función de un conjunto finito de "temas" ("topics") que detecta automáticamente.

Cada tema es un conjunto de números que cuantifica la probabilidad de cada una de las palabras del léxico o conjunto de palabras del corpus en ese tema.

Por ejemplo, si aparece un tema que corresponda al concepto "girar", tendrá probabilidades elevadas en palabras como girar, rotar, eje, rueda, revolución, etc.

Una vez obtenidas esas secuencias de números (vectores) que caracterizan cada tema, es posible, a su vez, caracterizar el corpus, cada texto del corpus e incluso un nuevo texto (que comparta léxico) con otro conjunto de números que cuantifican la probabilidad de que el corpus, el texto del corpus o el nuevo texto, trate de esos temas.

Por ejemplo, el texto de una patente de una nueva antena tendría probabilidades elevadas en temas como electromagnetismo, telecomunicaciones, conductores eléctricos, etc.

Lo interesante es que este vector de probabilidades de los temas tratados se convierte, por tanto, en una suerte de huella dactilar o firma del contenido temático del documento, que lo caracteriza.

Esta caracterización del documento, además de automática, es mucho más expresiva que las clasificaciones, que asignan a cada documento un único tema. Por otro lado, una vez convertido el contenido temático de los textos en vectores de números, el ordenador ya se encuentra muy cómodo para realizar operaciones diversas sobre estos vectores.

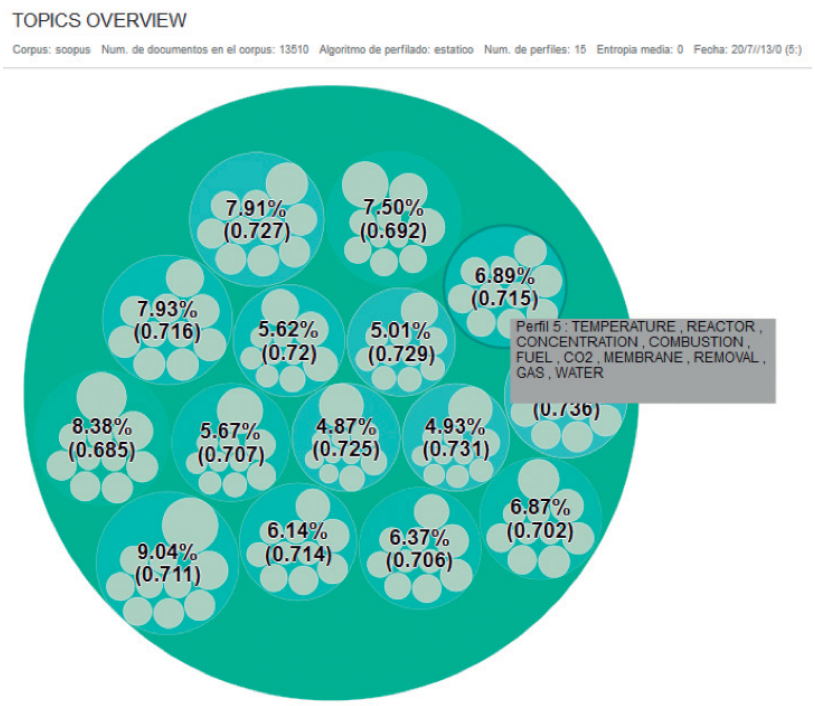


Figura 1. **Descomposición temática de un corpus mediante LDA (proyecto vigilancia sectorial).**

¿Para qué puede servir esta caracterización automática de los temas que trata un documento? Para muchas cosas, pero pongamos algunos ejemplos:

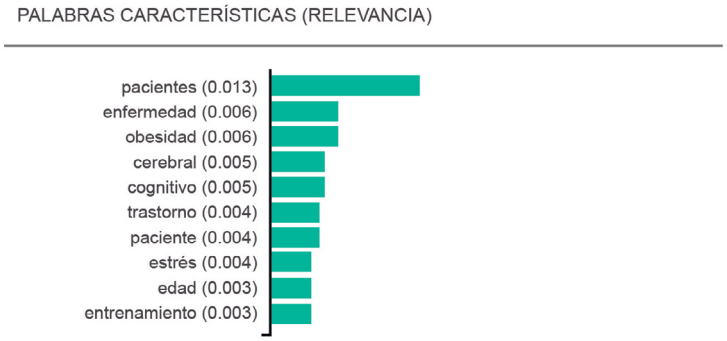
- Buscar en una colección de documentos, documentos similares por su contenido temático a uno dado (se trata de buscar vectores de números parecidos, cosa que un ordenador puede hacer muy rápidamente). Esto puede ayudar a evaluadores de ayudas públicas o de solicitudes de patentes a encontrar documentos similares a la solicitud estudiada. Esto es extremadamente útil cuando la colección de documentos en la que hay que buscar documentos similares es muy grande. En este proyecto se ha procesado, por ejemplo, el corpus de patentes TIC estadounidenses, que son casi un millón de documentos. Esta técnica, además, asigna a los mismos temas los sinónimos, de modo que permite detectar plagios sofisticados.

- Buscar documentos por su temática, es decir, de una forma más ajustada semánticamente que por palabras.

- Tener una visión de conjunto de los temas que trata un corpus documental. Así, por ejemplo, se ha caracterizado qué tecnologías tratan las solicitudes de ayudas a la SESIAD o las solicitudes de patentes TIC.

- Ver la evolución temporal de las temáticas. Por ejemplo la evolución temática en la innovación subvencionada por la SESIAD a lo largo de los años.

- Buscar solapamientos y sinergias en las políticas de ayudas públicas de diferentes departamentos, analizando las temáticas financiadas en unos y otros (por ejemplo, entre SESIAD, SEIDI, FECYT, CDTI).



DOCUMENTOS MEJOR CARACTERIZADOS POR ESTE PERFIL

%TOPIC	DOC ID	TITLE
100.00%	PSI2012-35352	ALTERACIONES NEUROENDOCRINAS E INMUNITARIAS EN RATONES CON DIFERENTES ESTRATEGIAS DE AFRONTAMIENTO DEL ESTRÉS SOCIAL CRÓNICO. EFECTO DEL TRATAMIENTO CON UN ANTAGONISTA CRH1
100.00%	PSI2008-00161	ESTUDIO DEL TRATAMIENTO CON UN AGONISTA SEROTONINERGICO DE ACCIÓN RÁPIDA SOBRE LOS EFECTOS CONDUCTUALE
100.00%	PSI2011-24762	EFECTOS DEL ESTRÉS SOCIAL EN EL CONDICIONAMIENTO DE LA PREFERENCIA DE LUGAR (CPL) INDUCIDO POR COCAÍNA Y MDMA (ÉXTASIS). IMPLICACIÓN DEL SISTEMA DOAMINERGICO Y GLUTAMATERGICO.

***Figura 2.* Ejemplo de tema ("topic") obtenido mediante LDA (proyecto vigilancia sectorial).**

Para catalizar este potencial, la Secretaría de Estado para la Sociedad de la Información y Agenda Digital (SESIAD) lidera el Plan de Impulso de las Tecnologías del Lenguaje, que se inserta en el marco de la Agenda Digital para España.

Este Plan tiene por objeto fomentar el desarrollo en España del sector del procesamiento del lenguaje natural, de la traducción automática o asistida por ordenador y de los sistemas conversacionales, y aprovechar estas novedosas capacidades para mejorar el servicio público.

El Plan de Impulso de las Tecnologías del Lenguaje, aprobado el 7 de octubre de 2015, se ha diseñado en apoyo a un sector cuyo diagnóstico de situación se podría resumir en las siguientes ideas clave:

- **Alto potencial de crecimiento y desarrollo:** Las tecnologías del lenguaje representan un sector emergente transversal vinculado a la innovación con capacidad para promover crecimiento, competitividad y empleo de calidad.

- **Oportunidad única:** Su desarrollo industrial, paralelo al desarrollo tecnológico, es imparable, y dada la importancia del idioma español en el mundo, la oportunidad es única para que ocupemos un espacio que sin duda alguien ocupará.

- **Recursos disponibles pero dispersos:** España dispone de los recursos necesarios para acometer ese desarrollo, pero para aprovechar la oportunidad es imprescindible impulsar y coordinar las actuaciones pertinentes desde la Administración General del Estado en coordinación con las Comunidades Autónomas y en colaboración con Iberoamérica y la Unión Europea.

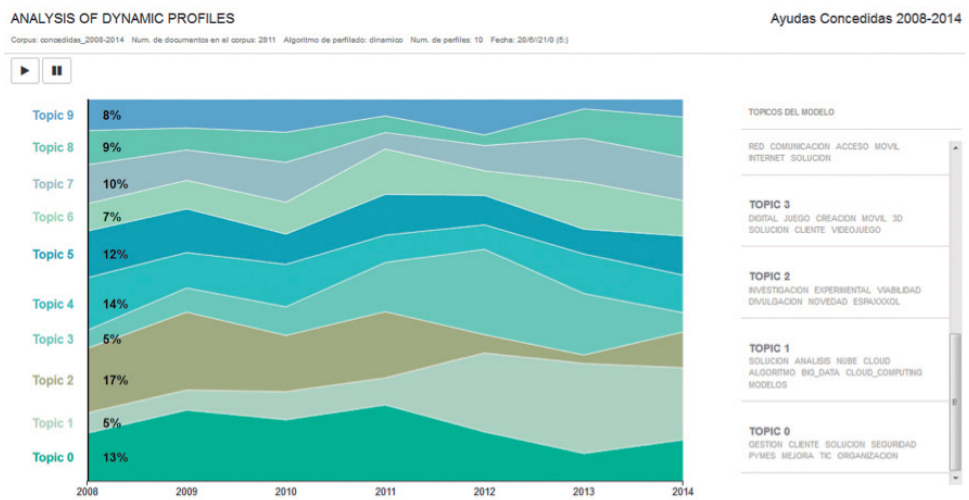


***Figura 3.* Evolución temporal de los temas ("topics") (proyecto vigilancia sectorial).**

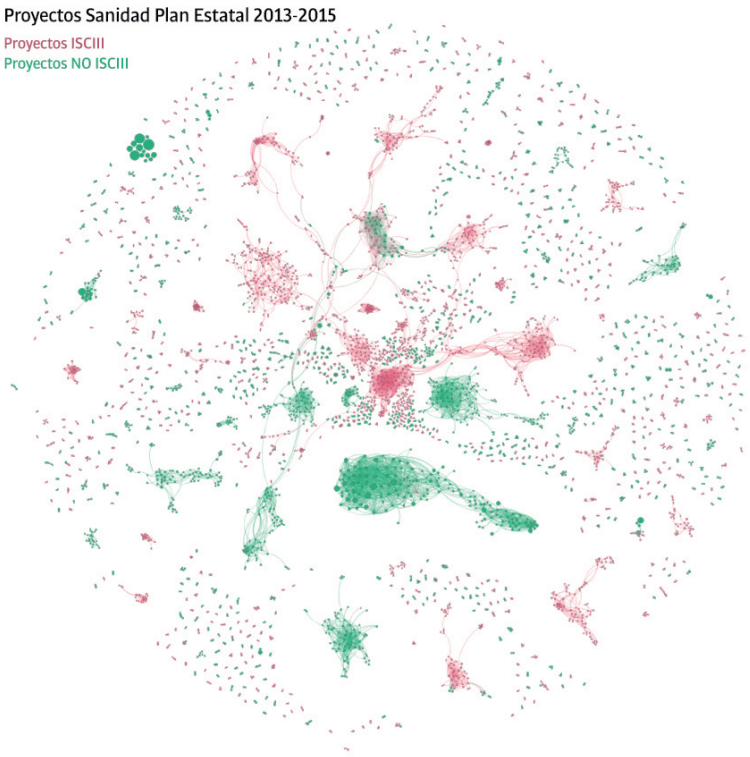


Figura 4. **Proyectos de sanidad solicitados agrupados por temática (proyecto vigilancia sectorial).**

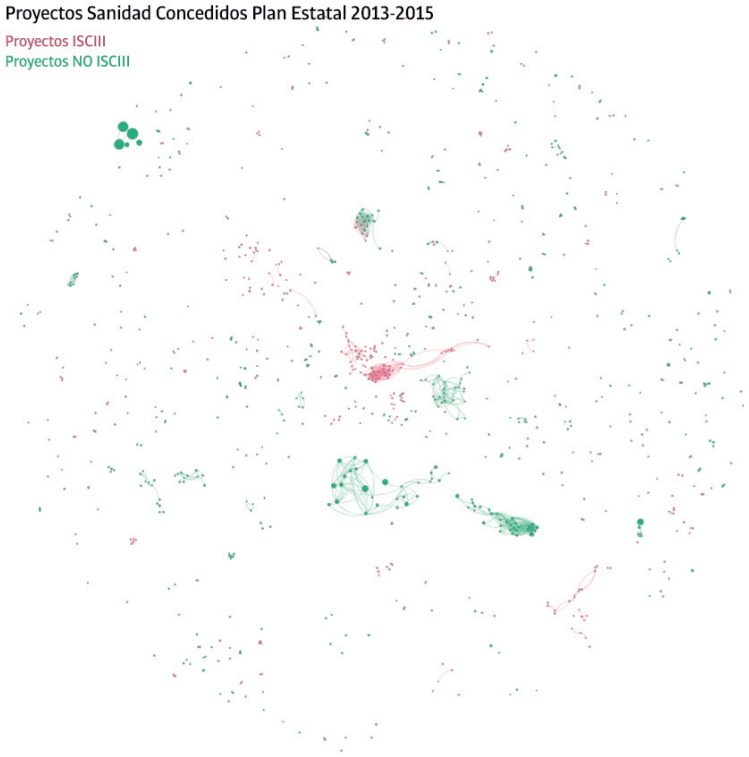


Figura 5. **Proyectos de sanidad concedidos agrupados por temática (proyecto vigilancia sectorial).**

FORTALEZAS	DEBILIDADES
• **Alto nivel investigador.** • **Buena gobernanza del idioma español** (RAE, ASALE). • **Experiencia en multilingüismo** por las lenguas cooficiales.	• Sector compuesto por **empresas demasiado pequeñas** como para competir en el mercado internacional o completar la cadena de valor en España. • **Transferencia insuficiente** del sector investigador a la industria.
OPORTUNIDADES	**AMENAZAS**
• **Importancia mundial el español.** • **Mercado en fuerte crecimiento** asociado a la innovación y al desarrollo. • **Sectores con gran potencial** (sanidad, turismo, educación, etc.). • **Reutilización de la información del Sector Público** (RISP) .	• **Pérdida de competitividad** económica e industrial de España e Iberoamérica. • **Subdesarrollo digital** del español y extinción digital de las lenguas cooficiales. • **Fuga de investigadores y profesionales** y deterioro del sector investigador español.

Análisis DAFO del sector español de tecnologías del lenguaje (*Fuente: Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España*).

El Plan de Impulso de las Tecnologías del Lenguaje

(www.plantl.es)

Se articula en cuatro ejes:

Eje 1: Desarrollo de infraestructuras lingüísticas

El objetivo de este eje del Plan es impulsar la industria del procesamiento del lenguaje natural en español y lenguas cooficiales poniendo a su disposición infraestructuras lingüísticas de propósito general. La existencia y disponibilidad de estos recursos permitirá a la industria española situarse en la frontera de la innovación, evitando la replicación de esfuerzos, logrando economías de escala y completando cadenas de valor en España.

Se pretende reducir la distancia en cantidad, calidad y disponibilidad, con las existentes en inglés.

Como toda infraestructura pública, las infraestructuras lingüísticas han de servir para mejorar procesos de terceros y han de garantizar acceso e interoperabilidad.

Eje 2: Impulso de la industria de las tecnologías del lenguaje

Las medidas de este eje se centran en las actuaciones de visibilidad, formación, transferencia del conocimiento e internacionalización.

Eje 3: La Administración Pública como impulsora de la Industria del Lenguaje

En este eje se buscan dos objetivos:

• Mejorar la calidad y capacidad del servicio público incorporando las tecnologías de procesamiento de lenguaje natural y de la traducción automática; actuando, a su vez, como tractor de la demanda.

• Impulsar la industria del procesamiento del lenguaje natural en español y lenguas cooficiales poniendo a su disposición recursos lingüísticos generados a partir de la información del sector público.

Para el primero de los objetivos, se propone la creación de sendas plataformas comunes de procesamiento de lenguaje natural y de traducción automática para las Administraciones Públicas, siguiendo el criterio de simplificar, lograr sinergias y aplicar la economía de escala en la puesta en marcha de nuevas capacidades y servicios basados en estas tecnologías.

Para el segundo de los objetivos, se quiere apoyar la generación, estandarización y difusión de recursos lingüísticos creados en el contexto de la actividad de gestión pública propia de la Administración, aprovechando el marco la política de Reutilización de la Información del Sector Público (RISP).

Eje 4: Proyectos Faro

Se trata de desarrollar nuevos servicios públicos, o mejorar la capacidad y calidad de servicios públicos existentes y de procedimientos de las Administraciones Públicas, mediante la aplicación de las tecnologías del lenguaje.

Para ello, los proyectos Faro persiguen llevar a cabo actuaciones en servicios públicos concretos de alto impacto social (sanidad, justicia, turismo, vigilancia sectorial, etc.) que abarquen toda la cadena de valor y den lugar a productos y servicios acabados, con el fin de poner en valor las capacidades y beneficios de las tecnologías del procesamiento del lenguaje natural y la traducción automática.

Estas actuaciones servirán también para guiar el diseño y dar aplicación inmediata a las medidas horizontales del plan, como el desarrollo de infraestructuras lingüísticas, y, muy

especialmente, a las plataformas comunes de procesamiento de lenguaje natural y traducción automática de las Administraciones Públicas.

Entre los instrumentos para implementar el Plan de Impulso de las Tecnologías del Lenguaje queremos llamar la atención sobre la Compra Pública de Innovación (CPI).

Como hemos señalado, para alcanzar sus objetivos, el Plan de Impulso de las Tecnologías del Lenguaje pretende llevar a la industria española a la frontera de innovación para hacerla competitiva a escala global, a la vez que se aprovechan estas novedosas capacidades para mejorar sustantivamente el servicio público.

Para ello hay que superar la paradoja por la cual el proveedor no invierte en productos innovadores, que requieren previamente una inversión en I+D+i, por falta de una demanda clara; y a su vez, el comprador no demanda productos innovadores porque no hay oferta disponible, adecuada y económica para los retos que tiene pendientes.

La falta de certidumbre sobre la demanda futura de la inversión en I+D+i es uno de los problemas que frena estos procesos, por ello la Administración, mediante su papel como comprador público, puede ofrecer esa necesaria certidumbre y tener un efecto tractor decisivo para el desarrollo de la industria.

La CPI se emplea cuando la Administración tiene necesidades que el mercado no cubre. Con la CPI se pretende que el sector industrial desarrolle productos o servicios para cubrir esas necesidades.

Como se trata de desarrollar productos o servicios nuevos es preciso conocer la capacidad del sector industrial para comprobar que el reto de innovación que se plantee, por un lado, no esté ya resuelto por el mercado, y por otro, sea factible desarrollarlo.

Precisamente por ello, en el marco de la CPI, se aconseja la utilización de la consulta preliminar al mercado. Se trata deun instrumento con el que la Administración busca obtener información para poder realizar una descripción funcional del producto o servicio innovador que necesita, con mayores probabilidades de éxito en el ulterior procedimiento de compra pública.

“Las tecnologías del lenguaje pueden mejorar servicios críticos para el ciudadano como sanidad, justicia o educación, y generar además importantes ahorros, como por ejemplo en la detección de fraude.”

Mediante una resolución del Secretario de Estado para la Sociedad de la Información y la Agenda Digital de 20 de abril de este año, se ha establecido el procedimiento de consulta preliminar al mercado mediante el cual se irán publicando los retos de innovación tecnológica que se van planteando en el marco del Plan (puede encontrar más información al respecto en *www.plantl.es*).

Por otro lado, con el fin de agregar demanda, perfilar mejor los retos de innovación e identificar otros ámbitos en los que se pueda conjugar la mejora del servicio público con el desarrollo de innovación en el ámbito de las tecnologías del lenguaje, se desarrolla una consulta simultánea dirigida al Sector Público que hemos denominado Expresión de Interés.

Queremos insistir en el papel clave de las Administraciones Públicas en el Plan de Impulso de las Tecnologías del Lenguaje.

Por un lado, las tecnologías del lenguaje pueden mejorar servicios críticos para el ciudadano como sanidad, justicia o educación, y generar además importantes ahorros, como por ejemplo en la detección de fraude.

Por otro, las Administraciones Públicas generan enormes cantidades de información textual en formato electrónico, buena parte de las cuáles se puede convertir en combustible para la industria de las tecnologías del lenguaje al amparo de RISP.

Y por último, la Administración puede actuar de tractor del sector español por su capacidad de liderazgo y generación de servicios sostenidos en el tiempo.

Por ello, queremos aprovechar la oportunidad de escribir en este medio para animar a los lectores a reflexionar sobre las oportunidades que ofrecen las tecnologías del lenguaje para mejorar la función pública en sus ámbitos de competencia y, en su caso, unirse al proceso de definición de retos de innovación tecnológica por el cauce habilitado al efecto en *www.plantl.es* o, sencillamente, contactado directamente con los autores de este artículo. ✱