GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA
SECRETARÍA DE ESTADO PARA LA SOCIEDAD DE LA INFORMACIÓN Y LA AGENDA DIGITAL

ontsi observatorio nacional de las telecomunicaciones y de la SI

red.es

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

ESTUDIO DE CARACTERIZACIÓN DEL SECTOR DE TECNOLOGÍAS DEL LENGUAJE EN ESPAÑA

Plan de Impulso de las Tecnologías del Lenguaje

acap asesores y consultores en administraciones públicas

oesia

Mayo 2018

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para la Sociedad de la Información y la Agenda Digital y Red.es, que no comparten necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

Índice

Resumen ejecutivo

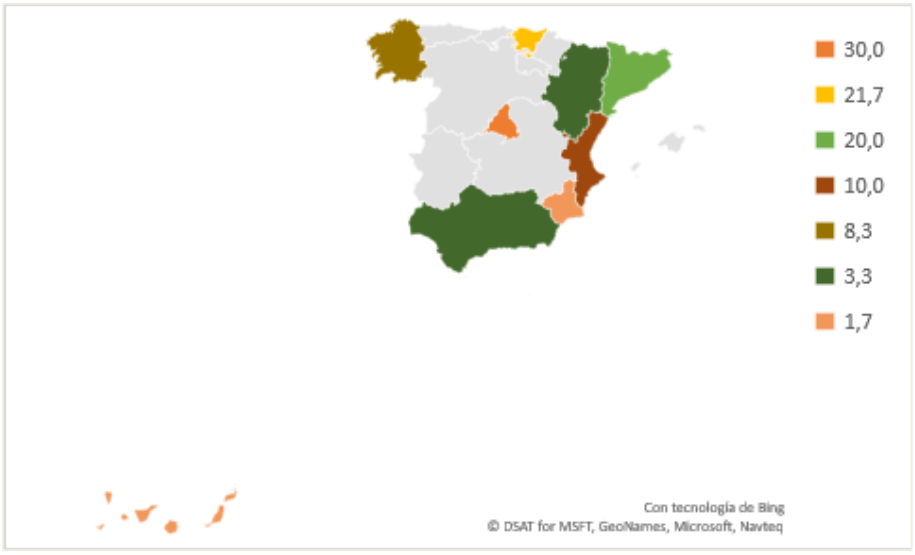
CARACTERIZACIÓN DE LAS EMPRESAS

El perfil de los agentes del sector

- Agentes identificados

Tipos de agentes	Identificación			
	Del sector de tecnologías del lenguaje	No se dedican al sector	Extinguidas, liquidadas, ilocalizables	Total
Empresas y asociaciones	127	69	19	215
Centros de investigación	63	8	0	71
Total	190	77	19	286

- La mayoría de los agentes del sector se concentran en la **Comunidad de Madrid** (30%), **País Vasco** (21,7%), **Cataluña** (20%) y **Comunidad Valenciana** (10%).



- La mayoría de los agentes tienen **menos de 30 años de antigüedad** (85,7%), y llevan entre 11 y 20 años dedicándose al sector.

Esto mostraría una actividad de tecnologías del lenguaje madura en lo que se refiere a la antigüedad de la actividad.

- La mayoría de los agentes desarrolla la actividad TL **combinada con otras líneas de negocio** (68,3%): **las empresas del sector desarrollan la actividad TL combinada con otras actividades relacionadas con el sector TIC** (31,4% "Otros servicios relacionados con las tecnologías de la

información y la informática", 17,1% "Actividades de programación informática" y "Actividades de traducción e interpretación"); **los centros de investigación coordinan la investigación con la actividad docente** ("Educación universitaria" 36%, "Otros servicios relacionados con las tecnologías de la información y la informática" 16% y "Actividades de programación informática" 12%).

Se ha detectado un problema de continuidad de la oferta de resultados de los centros de investigación en el mercado, ya que la mayoría de los centros (85,8%) no ha creado una spin-off, por lo que se quedan en la fase de investigación preindustrial y encuentran dificultades para dar el salto comercial hacia la fase industrial en el desarrollo de aplicaciones.

Personal ocupado del sector

- Poco más del 60% del total de los agentes del sector consultados tienen **menos de 49 empleados**.

Se ha identificado un sector conformado fundamentalmente por microempresas y pequeñas empresas.

- El 73% de los **agentes no especializados** consultados manifestó que el volumen de empleados de su empresa u organización vinculados a la actividad del sector se encuentra **entre los 5 y los 20 empleados**.
- La mayoría de los empleados del sector de los agentes consultados **son titulados superiores, doctores o ingenieros**, un 70,9%.
- Por otra parte, se ha **detectado una brecha de género** del 16,2%.
- El 75% de los agentes consultados afirmó **haber contratado personal durante el año 2017**.

Esto mostraría que el negocio de tecnologías del lenguaje está aumentando en las empresas y los centros de investigación.

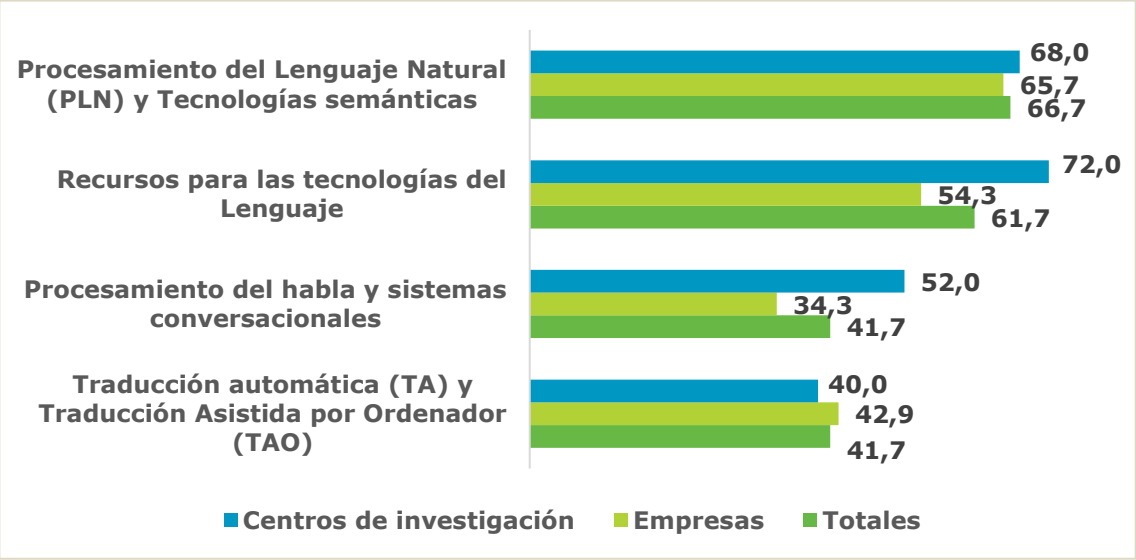
- Un 28,3% de los agentes expresaron **haber pensado en contratar personal y haber tenido dificultad para encontrar los técnicos adecuados** y un 21,6% aludieron a **no haber tenido los recursos económicos suficientes**.

A pesar de haber contratado personal, la mayoría de los agentes consultados señalan que existen dificultades del lenguaje por un problema de formación. La especificidad del sector reside en la necesidad de encontrar perfiles mixtos con conocimientos en tecnologías y conocimientos en lenguaje e idiomas.

Oferta de soluciones de tecnologías del lenguaje

- En general, la mayoría de los agentes del sector expresaron **comercializar con 2 o más tipos de productos y servicios de tecnologías del lenguaje** (traducción automática, procesamiento del habla y sistemas conversacionales, procesamiento del lenguaje natural y las tecnologías semánticas y recursos para las tecnologías del lenguaje), lo que indicaría que las actividades del sector de tecnologías del lenguaje **están de alguna forma interconectadas y no existe una gran especialización por tipo de producto o servicio que se comercializa entre los agentes consultados.**

Se ha identificado una actividad de tecnologías del lenguaje horizontal que utiliza herramientas básicas de todo tipo, tanto pertenecientes a la traducción automática, como a sistemas conversacionales, procesamiento del lenguaje natural o recursos para las tecnologías del lenguaje, para desarrollar finalmente las aplicaciones con orientación comercial.



- El 68% de los agentes que comercializan o investigan productos o servicios de Traducción automática (TA) y Traducción asistida por ordenador (TAO), **desarrollan motores de traducción automática**. En el caso de las empresas consultadas la especialización en este tipo de productos o servicios es todavía mayor (un 86,7%).
- En lo que respecta a las soluciones específicas que pertenecen a la categoría de procesamiento del habla y sistemas conversacionales, la actividad se encuentra muy repartida entre los **sistemas de diálogo o asistentes conversacionales** (60%), **el speech-to-text/text-to-speech** (60%) y **otras aplicaciones de procesamiento del habla** (56%).
- Respecto a las tareas relacionadas con el procesamiento del lenguaje natural y las tecnologías semánticas, la mayoría de los agentes consultados comercializa o desarrolla tareas de

preprocesamiento, tareas morfosintácticas, tareas sintácticas, tareas semánticas, extracción de terminología u otras tareas (reconocimiento de entidades nombradas y su clasificación, entity linking, extracción de relaciones, asistencia a la redacción, sistemas de asistencia a la pronunciación, polaridad, etc). La actividad que **menos desarrollan los agentes del sector es la relacionada con tareas programáticas y discursivas** con un 25%.

- El 64,9% de los agentes consultados que comercializan recursos para las tecnologías del lenguaje, producen **recursos lingüísticos, incluyendo corpus monolingües y bilingües**.
- La mayoría de las empresas y los centros de investigación consultados (86,7%) afirmaron orientar su actividad a la **lengua castellana y variables del castellano**. En un porcentaje algo menor (76,7%) los agentes del sector manifestaron orientar su actividad hacia **alguna lengua internacional** y, por último, algo más de la mitad orienta su actividad hacia alguna lengua cooficial.

El sector de las tecnologías del lenguaje es multilingüe, la mayoría de los agentes consultados orientan la actividad de su negocio o investigación en, al menos, tres lenguas.

- Todos los agentes del sector consultados dirigen su actividad en la lengua **inglesa**, lo que muestra el **gran predominio de este idioma en el mercado**. Las **empresas dirigen en mayor proporción su actividad en lenguas de países próximos** (francés 52%, portugués 43%, italiano y alemán 37%), **que los centros de investigación**, más orientados hacia la lengua inglesa.
- La lengua cooficial hacia la que más dirigen su actividad los agentes identificados es el **catalán** con un 73,5%.

El multilingüismo del estado español ha permitido que el sector de las tecnologías del lenguaje avance. La lengua castellana se posiciona a niveles similares que otras lenguas hegemónicas, a nivel de recursos y herramientas, aunque el inglés es la lengua predominante desde el punto de vista de soluciones de investigación y comercialización.

- La situación de la lengua castellana a nivel internacional mejora respecto a su situación a nivel europeo gracias a la **introducción del mercado latinoamericano**.

Volumen de ventas del sector

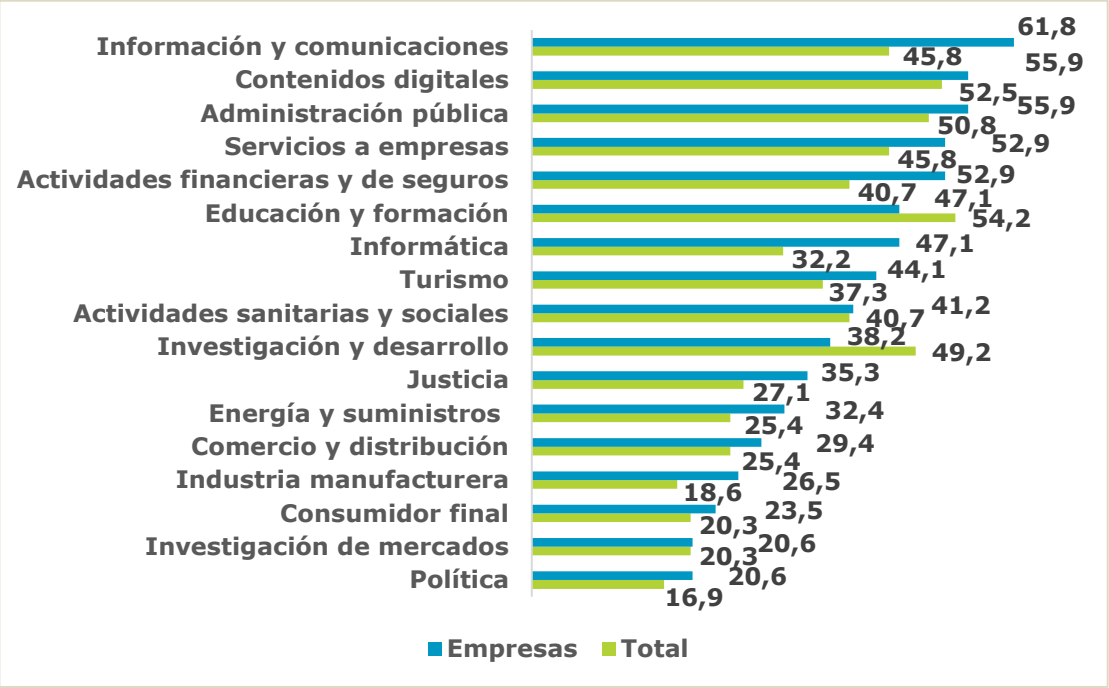
- Desde la perspectiva de las ventas se ha detectado un **sector en auge en la medida en que cerca de la mitad de los agentes consultados** (54,2%) **aumentaron su volumen de clientes en 2017**.

- El volumen de facturación que se corresponde con las actividades de tecnologías del lenguaje **se situó en 2016 alrededor de los 205 millones de euros**.

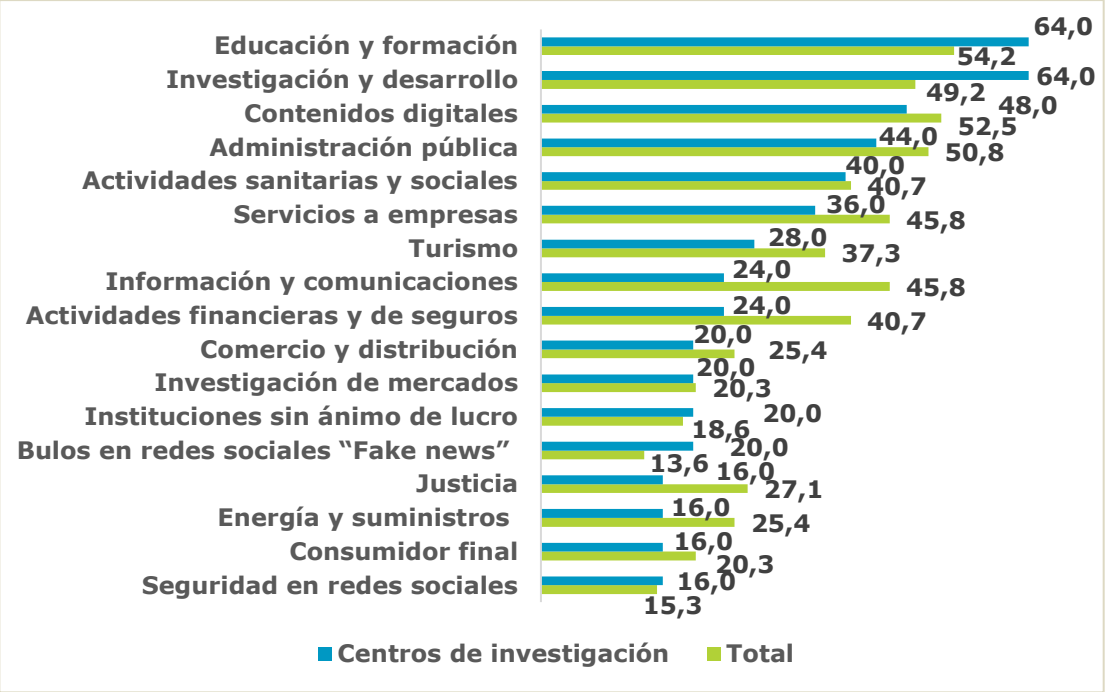
La actividad de tecnologías del lenguaje genera evidentes beneficios, lo que refuerza la idea de diferenciar este subsector en el ámbito de las TIC, mediante políticas y programas específicos dirigidos a su desarrollo.

Destino funcional de las ventas del sector

- Las empresas del sector dirigen sus ventas **hacia más de 11 sectores de actividad distintos**, lo que indica **la transversalidad de la aplicación de las tecnologías del lenguaje a todo tipo de sectores de actividad**, porque no se trata de tecnologías finalistas que vayan dirigidas a usos específicos. Las empresas dirigen sus ventas a **sectores con grandes volúmenes de consultas de usuarios finales.**



- Los centros de investigación están más orientados, de nuevo, a la investigación y la educación.



- En lo que respecta a la internacionalización del sector, el 52,5% de los agentes entrevistados ha indicado que **no exporta productos o servicios relacionados con las tecnologías del lenguaje a otros países.**
- El 58,8% de las empresas exporta productos o servicios a otros países, mientas **tan solo el 32% de los centros de investigación realiza exportaciones**.
- El ámbito geográfico al que van dirigidas en mayor proporción las ventas de los agentes consultados del sector es la **Unión Europea**, con un 82,1% de las ventas. Destaca que el 44% de los agentes comercializan en la Unión Europea en castellano, **lo que podría indicar el peso que tiene nuestra lengua en el ámbito europeo**.
- En el caso de las empresas consultadas del sector, el 55% dirigen sus ventas a Norteamérica y el 50% a Latinoamérica, mientras **tan solo el 25% de los centros de investigación dirigen sus ventas a estas áreas geográficas**.

Factores clave internacionalización: visibilidad en los mercados exteriores acompañada de presencia web; las empresas necesitan grandes cantidades de recursos lingüísticos para entrenar sus sistemas en otro idioma; los centros de investigación apuntan a la necesidad de colaborar en proyectos comunes al sector de tecnologías del lenguaje.

EL MODELO DE NEGOCIO

Cadena de valor

- La cadena de valor del sector de tecnologías del lenguaje podría plantearse en dos fases: una primera fase preindustrial, que va **desde el tratamiento de la materia prima hasta la aplicación de herramientas o componentes básicos a la información normalizada**, y una segunda fase industrial, **donde se desarrollan las aplicaciones que se comercializan en productos o servicios**, listas para ser consumidas por ciudadanos, empresas, administraciones, organizaciones o asociaciones.
- La solución comercializada está dirigida a dos tipos de clientes, por un lado, a un **cliente directo o final, es aquel que ha comprado el producto o servicio de tecnologías del lenguaje**, pudiendo ser una empresa, una administración, una institución, organización o asociación, y por otro lado, **a un cliente indirecto, que es el cliente que se beneficia de manera implícita de la tecnología del lenguaje aplicada**, este tipo de clientes suele ser la sociedad en su conjunto o los individuos particulares.

Modelo de ingresos

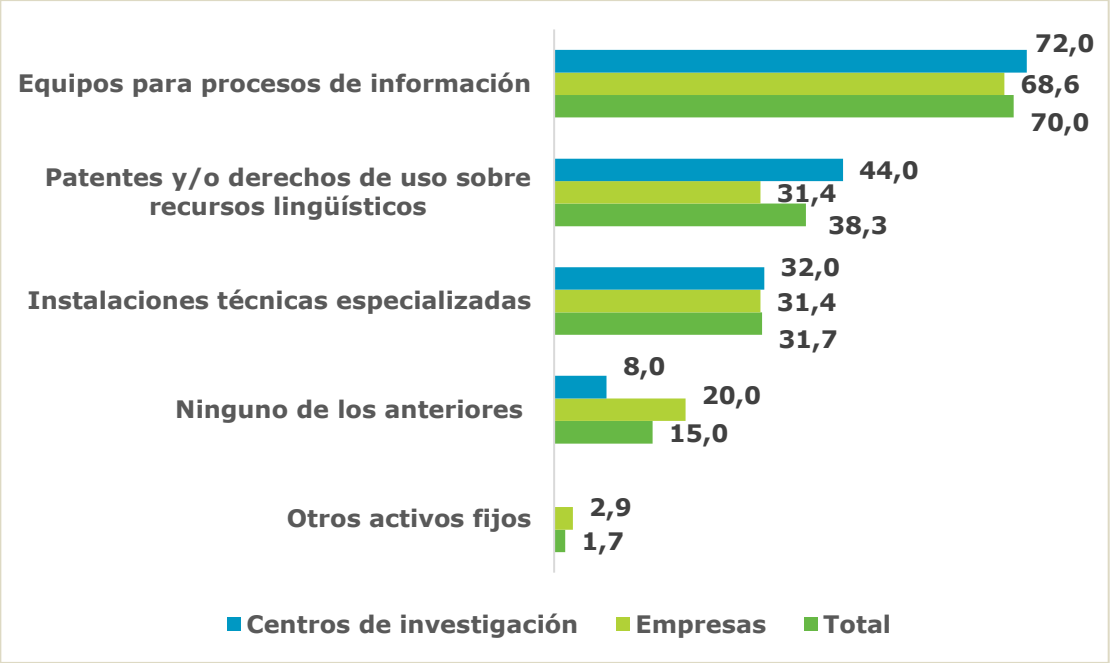
- El modelo de ingresos de las **empresas** del sector se basa fundamentalmente en la **venta de servicios profesionales para el cliente final por trabajo realizado.**

- El modelo de ingresos de los **centros de investigación** del sector se basa fundamentalmente en el **acceso a subvenciones de I+D+i.**

- El **software de código abierto** es un **instrumento que permite a los centros y universidades investigar y desarrollar soluciones TL**, a las que no tendrían acceso de otra forma por falta de recursos para integrar componentes de productos en el desarrollo de sus aplicaciones.

Modelo de producción

- Un 20% de las empresas consultadas **no mantienen activos fijos**, lo que podría indicar que tienen un **modelo de producción más orientado a ofrecer servicios de tecnologías del lenguaje para los que no precisan este tipo de activos**.



- Poco más de la mitad de los agentes consultados (55%) **no ha subcontratado servicios relacionados con las tecnologías del lenguaje a terceros**, aumentando hasta un 60% si atendemos a las empresas.

Estos datos podrían mostrar falta de relaciones de cooperación entre empresas.
Por otro lado, al tratarse de empresas que en su gran mayoría comercializan con soluciones de tecnologías del lenguaje de todo tipo, es decir, bastante heterogéneas en el tipo de soluciones que desarrollan, podrían no necesitar subcontratar ningún producto o servicio específico.

Investigación e innovación en el sector

- La mayoría de las empresas consultadas (61,8%) manifestó que tienen un departamento de I+D+i para apoyar el desarrollo de soluciones de su negocio, lo que **podría indicar el grado de innovación que implica el desarrollo de soluciones de tecnologías del lenguaje**.
- El **índice de inversión en I+D+i** de las empresas consultadas **representa el 6,3% de su facturación total**. Si atendemos al volumen de inversión que se corresponde con actividades del sector de tecnologías del lenguaje, **representa el 10,1% del volumen de su facturación que se corresponde con la actividad**.

La mayoría de las empresas considera que la inversión en I+D+i en el sector es escasa, tanto por parte de las empresas como por parte la administración.

- Respecto a las redes de conocimiento en las que participan los agentes consultados del sector, destaca la participación en **asociaciones especializadas** (40,7%). En segundo lugar, el 35,6% de los agentes consultados participa en **clústeres**.
- La mayoría de los **centros de investigación** (84%) **ha realizado publicaciones científicas**, **mientras las empresas están más enfocadas a la participación en líneas de investigación para el desarrollo de su negocio** (60,6%). La línea de investigación más desarrollada por los agentes es la **minería de datos (análisis de sentimiento)** con un 60%.
- En general, las empresas y centros de investigación afirmaron en su mayoría colaborar entre ellas, ya sea por medio de acuerdos entre empresas, universidades, asociaciones o centros tecnológicos. **Los acuerdos entre empresas se dan orientados a la comercialización concreta de productos o en el desarrollo de nuevas tecnologías**.

Modelo de "competición y cooperación": por un lado, entre empresas que se dedican a la misma actividad en el desarrollo de productos en la fase industrial, y por otro lado entre empresas que se dedican a distinta actividad en el desarrollo de soluciones donde se aporta el valor añadido de la colaboración (innovación abierta).

- Los **centros de investigación** expresaron que c**olaboran entre ellos en proyectos comunes o en el desarrollo de líneas de investigación**, y colaboran con las empresas **actuando como departamento de I+D+i**, como consultores o desarrolladores, **cediéndoles personal formado o realizando asesorías tecnológicas a las empresas**.

TENDENCIAS Y BARRERAS DEL SECTOR

Barreras del sector

Las empresas y centros de investigación apuntan hacia la falta de datos o corpus para entrenar los sistemas de tecnologías del lenguaje como principal barrera a la que se enfrenta el desarrollo del sector.

- La apertura de datos se enfrenta a **problemas legales relacionados con la Ley General de Protección de Datos**, existe determinada información más complicada de tratar y de abrir.
- Otro desafío al que se enfrenta la apertura de datos es **la interoperabilidad de los datos y los formatos en los que se pongan a disposición de los agentes.**

Los agentes señalaron la de falta de formación de profesionales del sector como barrera para el desarrollo del sector. Esta falta de formación está vinculada a la multidisciplinariedad asociada a la actividad de tecnologías del lenguaje, que precisa de perfiles mixtos formados por lingüistas y técnicos.

Las empresas y los centros de investigación del sector expresaron que existe un obstáculo en su desarrollo relacionado con la naturaleza de su tamaño. El sector de tecnologías del lenguaje está conformado por empresas y centros de investigación pequeños, lo que dificulta ganar cuota de mercado frente a competidores de otros países.

- Algunas empresas señalan que las tecnologías del lenguaje corren el riesgo de convertirse en una *commodity* en la medida en que se trata de una actividad transversal a la mayoría de los sectores productivos y no aparece como elemento independiente que aporta valor.

Tendencias y principales oportunidades del sector

La oportunidad se encuentra en todas las actividades económicas y administrativas que puedan integrar las nuevas tecnologías, especialmente en aquellos sectores que requieran una interacción con el usuario final, dado que los usuarios demandan cada vez más una interacción más natural e inmediata.

Destaca el sector Big Data en una economía mundial que se encuentra en un proceso de transformación digital marcado por la necesidad de gestión de la ingente cantidad de datos y contenidos a través de Internet, redes sociales y medios digitalizados. En este sentido, el procesamiento de esa información y su puesta en valor depende en gran medida del análisis del lenguaje natural. Existe una oportunidad estratégica en la aplicación de las tecnologías del lenguaje natural a los procesos de lenguaje no estructurado, una minería de datos de valor.

Destaca el sector sanitario, concretamente la investigación biomédica, donde las tecnologías del lenguaje tienen tres aplicaciones detectadas como potenciales: el soporte a la decisión clínica, el enriquecimiento de colecciones de datos para la investigación y la codificación de diagnósticos.

El rol de la Administración

- Las empresas y centros de investigación conocen los programas y ayudas que ofrecen las administraciones públicas relacionados con el I+D+i, no obstante, **consideran que existen factores que inhiben la participación** en subvenciones y programas públicos, algunos de ellos son factores transversales a cualquier sector de actividad.
    - El **pequeño tamaño de las empresas**, que implica escasos recursos económicos y de personal para afrontar la carga burocrática que lleva asociada la presentación de un proyecto.
    - La **falta de seguimiento posterior de las subvenciones que se aprueban**, lo que impide medir el impacto real que tienen en las empresas o los centros de investigación beneficiarias.

- **Los pliegos están orientados a bienes tangibles**, no se tiene en cuenta las características y particularidades que tiene el desarrollo de software con respecto a la industria tangible, que es totalmente medible, cuantificable y comparable.

El eje IV: Proyectos Faro del plan de Impulso de Tecnologías del Lenguaje tiene como objetivo desarrollar proyectos emprendidos por las Administraciones Públicas de aplicación de las tecnologías del lenguaje en sectores estratégicos que pretenden servir de demostración de sus capacidades y beneficios, generar industria y generar recursos reutilizables en otros proyectos.

Estos proyectos impulsados por la SESIAD están dirigidos a los llamados sectores verticales y se instrumentalizarán a través de la creación de una oficina técnica general para cada uno de los sectores verticales. Las oficinas técnicas pretenden reunir a grupos de expertos o competentes que asesoren a cada administración en la creación de productos piloto que después puedan lanzarse al mercado en forma de compra pública innovadora, como una suerte de catalizador.

Los componentes de esos productos piloto han de ser totalmente interoperables, atómicos y modulares, de manera que se puedan dividir cada uno de los componentes del producto para su comercialización. La necesidad de atomizar los productos viene motivada por la última fase de intervención de la oficina técnica general, la evaluación de dichos productos sobre una base científica. Esta evaluación permitirá que la administración compruebe los componentes de producto que han funcionado y los que se podría mejorar.

1 Introducción

1.1 Objeto y alcance del estudio

El estudio de caracterización sobre el sector de tecnologías del lenguaje en España ha sido elaborado por ACAP y OESIA para el Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información (ONTSI) de Red.es y para la Secretaría de Estado para la Sociedad de la Información y la Agenda Digital (SESIAD).

El objeto del estudio es caracterizar a la industria y a los agentes que la conforman, conocer la situación actual y la evolución reciente de las características estructurales y económicas específicas de cada una de las actividades que componen el sector de las tecnologías del lenguaje en España.

El Plan de Impulso de las Tecnologías del Lenguaje[1] se ejecuta en el marco de la Agenda Digital para España y tiene como objetivo fomentar el desarrollo del procesamiento del lenguaje natural, los sistemas conversacionales y la traducción automática en lengua española y en lenguas cooficiales. Para ello, establece un conjunto de medidas encaminadas a aumentar el número, calidad y disponibilidad de las infraestructuras lingüísticas en español y lenguas cooficiales. En paralelo, trata de impulsar la industria del lenguaje fomentando la transferencia de conocimiento entre el sector investigador y la industria, e incorporando a la Administración como impulsora de este nuevo sector.

El plan se estructura en cinco ejes principales, entre los que se encuentra el Eje II: Impulso de la Industria de las Tecnologías del Lenguaje, que responde al objetivo de apoyar la transferencia de conocimiento entre el sector investigador y la industria, así como la internacionalización de las empresas e instituciones que componen el sector.

Los objetivos específicos del plan son los siguientes:

- Aumentar el número, calidad y disponibilidad de las infraestructuras lingüísticas en español y lenguas cooficiales.
- Impulsar la Industria del lenguaje fomentando la transferencia de conocimiento entre el sector investigador y la industria. Ayudar a la internacionalización de las empresas e instituciones que componen el sector. Mejorar la difusión de los proyectos actuales

[1] www.plantl.es

- Mejorar la calidad y capacidad del servicio público incorporando las tecnologías de procesamiento de lenguaje natural y de la traducción automática, actuando, además, como tractor de la demanda. Apoyar la generación, estandarización y difusión de recursos lingüísticos creados en el contexto de la actividad de gestión pública propia de la Administración.

1.2 Metodología

La metodología empleada en el estudio se basó en el método de triangulación, que combina la realización de entrevistas en profundidad, grupos de discusión y una encuesta online, todo ello apoyado sobre la revisión bibliográfica.

A nivel cualitativo se han mantenido 26 entrevistas en profundidad a empresas, centros de investigación y agentes de la administración pública en relación a diversas áreas temáticas del sector de las Tecnologías del Lenguaje.

Agentes	Entrevistas
Empresas	13
Centros de investigación	7
Administración pública	6

Además, se han desarrollado 4 grupos de discusión, 2 iniciales, uno con empresas del sector y otro con centros de investigación, y 2 grupos al finalizar el trabajo de campo cuantitativo, con empresas y centros de investigación del sector.

Periodo	Composición
Pre-encuesta	Un grupo de discusión con empresas Un grupo de discusión con centros de investigación
Post-encuesta	Dos grupos de discusión mixtos

Por otro lado, a nivel cuantitativo se ha realizado una encuesta online a empresas y centros de investigación del sector a nivel nacional, para obtener información general sobre la situación actual del sector de las Tecnologías del Lenguaje.

Agentes	Encuestas
Empresas	35
Centros de investigación	25
Total	60

1.3 Las tecnologías del lenguaje

Las tecnologías del lenguaje tienen un papel transformador en la economía, las administraciones, la ciudadanía, y en la forma en que los agentes involucrados en los procesos económicos, políticos, sociales y culturales interactúan y se comunican. Esta transformación se produce en cuatro sentidos:

1. **Las tecnologías del lenguaje transforman la interacción "Hombre-Máquina"** a partir de la innovación y la aplicación del conocimiento lingüístico a la tecnología, lo que conlleva una transformación en el modo de comunicarse de las empresas, los ciudadanos y las administraciones.
2. **La inteligencia artificial aplica las tecnologías del lenguaje para la creación de módulos capaces de procesar el lenguaje natural** y capaces de interactuar con las personas, lo que supone una transformación de los procesos productivos de las empresas y de la organización y el funcionamiento de las administraciones en tanto constituyen una industria habilitadora que participa en multitud de aplicaciones y dispositivos. Los avances tecnológicos en el procesamiento del lenguaje reportan grandes beneficios en áreas tan determinantes para la evolución de la sociedad como la investigación biomédica o la atención sociosanitaria, por lo que también suponen una transformación de la vida de las personas.
3. **Las tecnologías del lenguaje favorecen la creación de una sociedad multilingüe**. Desde una perspectiva económica, la aplicación de las tecnologías del lenguaje contribuye a superar las barreas lingüísticas que presenta la consecución de un mercado único digital en la Unión Europea, donde conviven hasta 24 lenguas oficiales y más de 60 lenguas nacionales o regionales, lo que podría representar un obstáculo para la libre circulación de productos y servicios, especialmente para las PYMES, componente esencial del mercado único digital. Cabe señalar que el 15% de las PYME europeas venden en línea (Cracking the language barrier federation, 2016), concretamente en el caso de España, tan solo el 6% de las empresas realizan ventas electrónicas a otros países de la Unión Europea (European Commission, 2017).

   Las tecnologías del lenguaje podrían contribuir a aumentar las ventas electrónicas a países con idiomas diferentes entre pymes, por lo que se convertiría en un sector importante para el impulso del crecimiento de las pymes españolas y para la mejora de su competitividad en una economía global.

   Además, las tecnologías del lenguaje representan una oportunidad de negocio en el mercado latinoamericano, en tanto el español es la segunda lengua con más hablantes nativos en el mundo y la tercera por número de hablantes (Núria & German, 2015) y la mayoría de las aplicaciones de tecnologías del lenguaje que se desarrollan están orientadas a procesar el inglés, por lo que el mercado latinoamericano se convierte en una oportunidad de mercado para las empresas españolas donde pueden desarrollar aplicaciones en español. Cabe mencionar como elemento de oportunidad de mercado del sector, la experiencia de las empresas españolas en la gestión del multilingüismo, dada la convivencia de cuatro lenguajes cooficiales que han permitido a las

empresas y los centros de investigación españoles ser punteros en el sector de las tecnologías del lenguaje.

Desde una perspectiva institucional, las tecnologías del lenguaje pueden ayudar a las administraciones a optimizar sus funciones, sus procesos de gestión y los servicios al ciudadano, así como a obtener información y conocimiento valioso de los datos que manejan.

4. **Las tecnologías del lenguaje son decisivas en el tratamiento de grandes cantidades de datos no estructurados** y se convierten en una herramienta esencial en el desarrollo del sector Big Data. La mayoría de estos datos están en el lenguaje natural, lo cual precisa de la aplicación de las tecnologías del lenguaje para su explotación.
   La clasificación de datos y el análisis de contenidos son tecnologías clave para sacar provecho a grandes cantidades de datos no estructurados de manera que se pueda impulsar su análisis: a través de la homogeneización de datos, el análisis semántico, el enriquecimiento y la reutilización de datos.

   Así pues, la economía de los datos requiere nuevos mecanismos innovadores que permitan aprovechar las cadenas de valor de los datos y los textos, su comprensión y su aplicación a los sectores productivos de la economía.

En resumen, las tecnologías del lenguaje se encuentran detrás de muchos productos digitales cotidianos: las comunicaciones móviles, los medios sociales, los asistentes inteligentes y las interfaces de voz, que están transformando la forma en que los ciudadanos, las empresas y las administraciones públicas interactúan en el mundo digital, por lo que se podría definir el sector de las tecnologías del lenguaje de la siguiente manera:

Las tecnologías del lenguaje son un conjunto de sistemas de software diseñados para manejar el lenguaje humano en todas sus formas, permiten analizar lenguaje escrito, hablado y facilitar su explotación en aplicaciones informáticas de uso en los sectores productivos de la economía.

La mayoría de las aplicaciones basadas en las tecnologías del lenguaje comparten un amplio y heterogéneo grupo de técnicas y herramientas básicas para el análisis y la producción de lenguajes, que podrían clasificarse en cuatro grandes tipos de actividades: la traducción automática y traducción asistida por ordenador, el procesamiento del habla y los sistemas conversacionales, el procesamiento del lenguaje natural y las tecnologías semánticas y los recursos para las tecnologías del lenguaje.

A continuación, se definen las soluciones que integran la actividad del sector:

1. **Traducción automática (TA) y Traducción Asistida por Ordenador (TAO)** (Berner, 2015)**:** las tecnologías relacionadas con la traducción automática y la traducción asistida son sistemas y herramientas que permiten traducir de un lenguaje natural a otro.

Al igual que la traducción de un idioma natural a otro hecha por los humanos, la traducción automática no consiste simplemente en sustituir palabras en un idioma por otro, sino en la aplicación de conocimientos lingüísticos complejos: morfología (cómo se construyen las palabras a partir de unidades de significado más pequeñas), sintaxis (estructura de las oraciones), semántica (significado) y pragmática (contexto).

Las soluciones de traducción automática y traducción asistida por ordenador utilizan las siguientes herramientas o componentes básicos:

Herramientas de traducción asistida: las herramientas de traducción asistida por ordenador son herramientas que guardan los segmentos traducidos y los segmentos originales como unidades de traducción distintos. Estas herramientas permiten crear y gestionar bases de datos, denominadas memorias de traducción, en las que se van almacenando las traducciones que los usuarios realizan.

Motores de traducción automática: son sistemas que permiten traducir automáticamente de una lengua origen a una lengua destino.

2. **Procesamiento del habla y sistemas conversacionales:** las tecnologías relacionadas con el procesamiento del habla y los sistemas conversacionales integran el procesado de la voz y la transformación de una secuencia de palabras reconocidas en forma de texto, y su reproducción en sistemas conversacionales. Normalmente incluye el reconocimiento del habla, la comprensión del lenguaje hablado, la conversión a texto, la síntesis de voz a partir de texto. Por otro lado, los sistemas de diálogo suelen incluir un componente de interacción, negociación y generación del habla.

Las soluciones relacionadas con el procesamiento de habla y los sistemas conversacionales utilizan las siguientes herramientas o componentes básicos:

Sistemas de diálogo o asistentes conversacionales (Pérez & Llorens, 2017): los asistentes conversacionales son sistemas informáticos que reciben como entrada frases del lenguaje natural expresadas de forma oral y generan como salida frases del lenguaje natural expresadas asimismo de

forma oral. La finalidad de estos sistemas es emular el comportamiento inteligente de un ser humano que realiza una tarea concreta.

En esta categoría se encuentran los chatbots, que son programas que simulan mantener una conversación con una persona al proveer respuestas automáticas a entradas realizadas por el usuario.

Speech-to-text/text-to-speech (European Commision, 2014)**:** los sistemas speech-to-text reconocen el lenguaje hablado y lo transcriben a texto. Los sistemas de text-to-speech parten de un texto y lo sintetizan en forman de lenguaje hablado.

3. **Procesamiento del Lenguaje Natural (PLN) y Tecnologías semánticas (basadas en análisis de texto)** (Palomar)**:** integran un conjunto de tecnologías que investigan y formulan mecanismos computacionalmente efectivos capaces de reconocer, analizar, comprender y generar el lenguaje.

El procesamiento del lenguaje natural puede clasificarse en cinco niveles de análisis: tareas de preprocesamiento, tareas morfosintácticas, tareas sintácticas, tareas semánticas y tareas pragmáticas y discursivas.

Tareas de preprocesamiento: las tareas de preprocesamiento integran tareas de tokenización (reconocer los elementos que aparecen en el texto: palabras, números, símbolos, signos de puntuación, etc.) y segmentación (reconocer las frases y párrafos), aplicadas a la identificación de un idioma.

Tareas morfosintácticas: implican el empleo de herramientas que analizan las palabras para extraer raíces, rasgos reflexivos, unidades léxicas compuestas y otros fenómenos.

Tareas sintácticas: implican el empleo de herramientas que analizan la estructura sintáctica de la frase mediante analizadores de base estadística o gramática de la lengua en cuestión a través del chunking (método para análisis superficial de frases de lenguaje natural en estructuras sintácticas parciales) y el análisis sintáctico.

Tareas semánticas: son herramientas semánticas que se basan en la extracción del significado al nivel de una palabra (léxico-semántico) o al nivel de una frase (sintáctico-semántico) y la resolución de ambigüedades léxicas y sintácticas a través de la anotación semántica y la anotación de roles semánticos.

Por ejemplo, se aplican herramientas de identificación de la acepción empleada (*Word sense desambiguation*) a cada elemento que pueda tener varias acepciones o herramientas de identificación, clasificación y desambiguación de entidades nombradas (*Named entity recognition and classification*), que incluyen: nombres propios-personas, lugares, organizaciones, expresiones numéricas-fechas, cantidades de dinero, etc.

Tareas pragmáticas y discursivas: las tareas pragmáticas y discursivas integran herramientas de análisis pragmático relacionadas con la anotación de marcadores del discurso y resolución de correferencias y herramientas de análisis pragmático relacionadas con el análisis del texto más allá de los límites de la frase.

Por ejemplo, se aplican herramientas para determinar los antecedentes referenciales para identificar cada una de las diferentes menciones al mismo objeto en el texto o herramientas de planificación del texto para generar texto que permita estructurar cada frase con el fin de expresar el significado adecuado.

4. **Recursos para las tecnologías del Lenguaje** (Listerri, 1999)**:** los recursos para las tecnologías del lenguaje son corpus, lexicones, bases de datos terminológicas y ontologías que se emplean para obtener el conocimiento necesario para el desarrollo de herramientas y sistemas del sector.

Corpus: Los corpus lingüísticos están constituidos por un conjunto de ejemplos reales de uso de una lengua. Integra un conjunto de textos almacenados en formato electrónico y agrupado, con el fin de estudiar una lengua o una determinada variedad lingüística. Su objetivo es constituirse en elementos de referencia para el estudio de una frase concreta o un cierto aspecto de una lengua. Son corpus textuales de carácter general el “Corpus de Referencia del Español Actual” (CREA), o el “Corpus Diacrónico del Español” (CORDE), ambos en desarrollo por la Real Academia española.

Lexicones: Los lexicones son instrumentos que integran la representación del conocimiento léxico asociado a una lengua, de forma que va más allá de un simple repositorio de palabras, dado que integra 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).

Bases de datos terminológicas: Una base de datos terminológica es una base de datos que permite el acceso a términos y su correspondencia en distintos idiomas. Está formada por los términos en el idioma de origen y las correspondientes equivalencias en uno o más idiomas de destino. Junto con los términos del idioma de origen se pueden guardar definiciones, frases de contexto, imágenes y otras muchas informaciones, relacionadas. Pueden estar vinculadas al uso de los términos en determinados entornos. A modo de ejemplo son bases de datos terminológicas: UNTERM (Base de datos multilingüe de las Naciones Unidas) FAOTERM (Base de datos multilingüe de la Organización de las Naciones Unidas para la Agricultura y la Alimentación- FAO), EUSKALTERM (Banco de datos terminológico público vasco), IATE (El Inter Active Terminology for Europe es un banco de datos terminológicos del Servicio de Traducción de la Comisión Europea, que contiene términos, abreviaturas, acrónimos y fraseología en las lenguas oficiales de la Unión Europea).

Ontologías: Una ontología está constituida por una lista de términos formales y una lista formal de definiciones/restricciones para esos términos. 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 considera como ontología a toda organización cognitiva que oscile desde la noción más simple de las taxonomías, pasando por los tesauros y modelos conceptuales hasta llegar a las teorías lógicas. Se entiende que una ontología define conceptos (significados) usados para describir y representar un área de conocimiento.

Para terminar, cabe señalar algunas de las aplicaciones que utilizan componentes de procesamiento del lenguaje natural:

- Inteligencia artificial: las aplicaciones inteligentes son amplias, pero entre las soluciones inteligentes que emplean PLN se encuentran los sistemas de clasificación o apoyo a la decisión.
- Minería de datos y Text Analytics que se basan principalmente en modelos de aprendizaje automático, pero suelen tener unos componentes de PLN para analizar los textos y suelen basarse en modelos de lenguaje calculados a través de modelos de aprendizaje automático y vectorización.
- Extracción de información: sistemas que encuentran y relacionan segmentos que representan información relevante de un conjunto de textos y descartan otras informaciones no relevantes.
- Recuperación de información: sistemas que, a partir de un conjunto de textos, proporcionan un subconjunto de texto que contienen información relevante (información solicitada por el

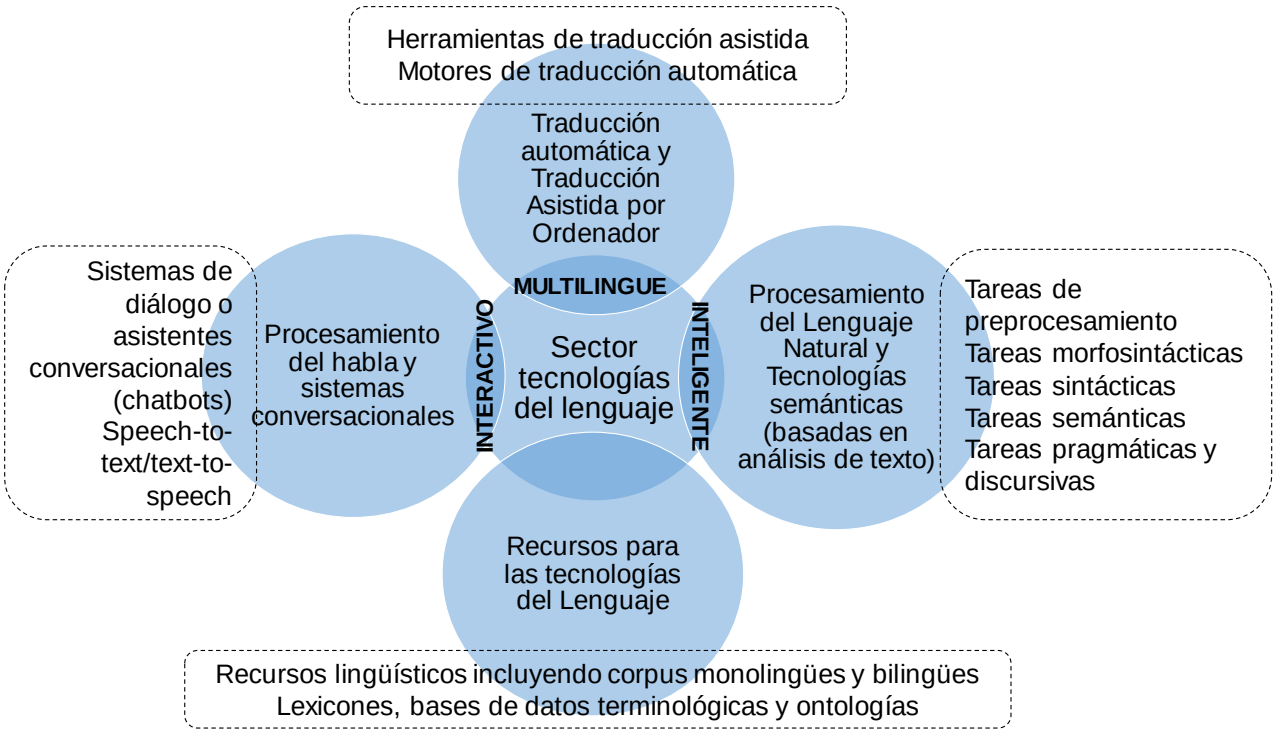
usuario). Aquí los métodos de indexación se pueden optimizar empleando componentes de PLN para describir el contenido del texto.

- Análisis de opiniones y sentimientos: herramientas que tratan de identificar la emoción asociada al texto (por ejemplo, aprobación o repulsa).
- Question answering. Los sistemas denominados "question answering" (QAS) o de búsqueda de respuestas, se relacionan generalmente con los motores de búsqueda, pero constituyen un tipo de recuperación de la información, más sofisticada, basada en la obtención de respuestas inteligentes a preguntas generadas en lenguaje natural.
- Resúmenes automáticos. Los resúmenes por extracción son procesos que permiten la obtención, el filtrado, la clasificación y la extracción de información, de documentos más amplios en función de distintos parámetros. Los resúmenes por abstracción utilizan técnicas más sofisticadas de tratamiento del lenguaje, ya que el resultado no consiste en determinadas oraciones entresacadas del texto original, sino en un documento de nueva redacción generado a partir del tratamiento de la información contenida primero.
- Text entailment, proceso denominado vinculación textual en castellano. Permite establecer relaciones direccionales entre dos expresiones en las que la segunda (denominada hipótesis vinculada) se puede deducir del significado de la primera expresión en el caso de una interpretación común de ésta.

A continuación, se ofrece de forma gráfica la **clasificación** que se ha realizado de las cuatro grandes tipologías de soluciones de tecnologías del lenguaje.

El sector de las tecnologías del lenguaje se basa en los recursos para las tecnologías del lenguaje para el desarrollo de aplicaciones, componentes, productos y soluciones interactivas, a través del procesamiento del habla y los sistemas conversacionales, multilingües, a través de la traducción automática y la traducción asistida por ordenador, e inteligentes, a través del procesamiento del lenguaje natural y tecnologías semánticas.

FIGURA 1. CLASIFICACIÓN SOLUCIONES TECNOLOGÍAS DEL LENGUAJE



2 Caracterización de los agentes del sector

2.1 El perfil de los agentes del sector

Número de agentes identificados en el censo

En el proceso de elaboración del censo se han identificado 127 empresas que se dedican a la actividad de tecnologías del lenguaje y 63 centros de investigación, un total de 190 agentes del sector de tecnologías del lenguaje conforman el censo.

TABLA 1: AGENTES IDENTIFICADOS EN EL CENSO

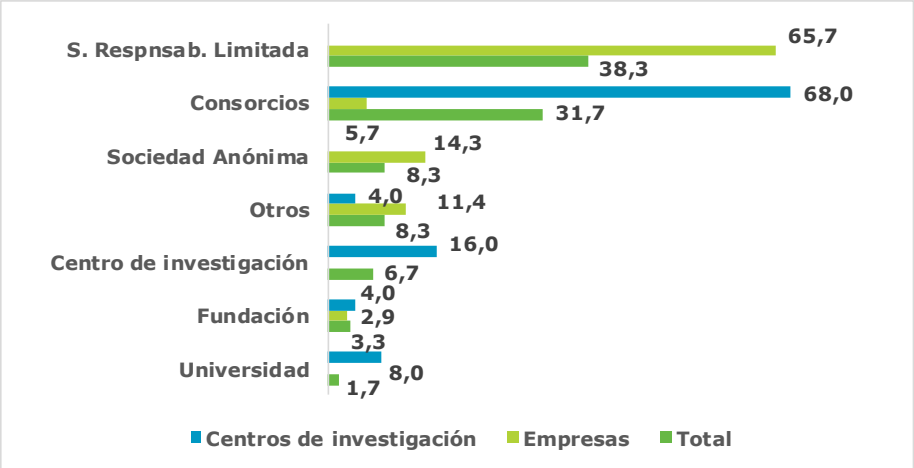
Tipos de agentes	Identificación			
	Del sector de tecnologías del lenguaje	No se dedican al sector	Extinguidas, liquidadas, ilocalizables	Total
Empresas y asociaciones	127	69	19	215
Centros de investigación	63	8	0	71
Total	190	77	19	286

Personalidad jurídica de los agentes del sector

Más de la mitad de los agentes consultados en la encuesta web manifestaron que el régimen jurídico de sus empresas o centros de investigación era el de **Sociedad de Responsabilidad Limitada** (38,3%) o **Consorcios** (31,7%).

Concretamente, el régimen jurídico más nombrado entre las empresas fue el de Sociedad de Responsabilidad Limitada con un 65,7% de las respuestas, mientras el 68% de los centros de investigación afirmaron que formaban parte de Consorcios.

FIGURA 2. FORMA JURÍDICA AGENTES DEL SECTOR %

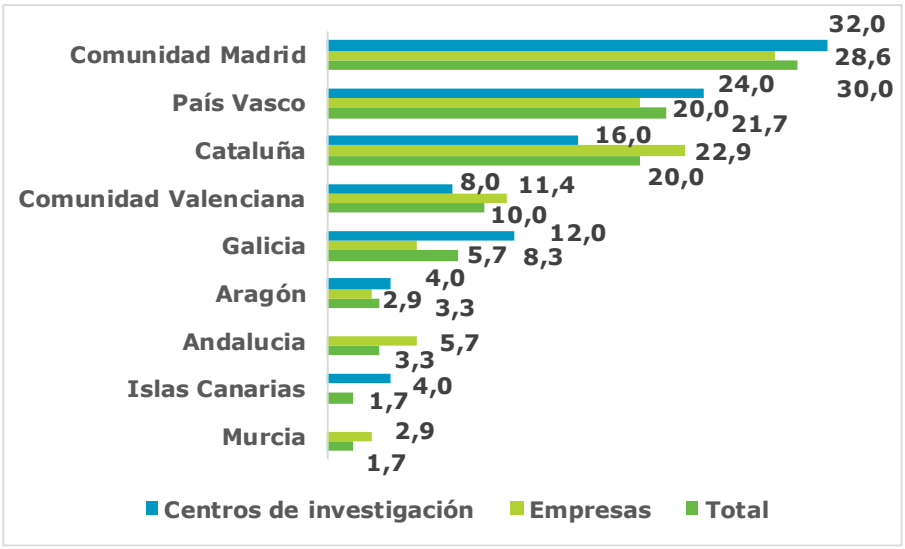


Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. F5

Ámbito geográfico del sector

La mayoría de los agentes consultados del sector se concentran en la **Comunidad de Madrid** (30%), **País Vasco** (21,7%), **Cataluña** (20%) y **Comunidad Valenciana** (10%).

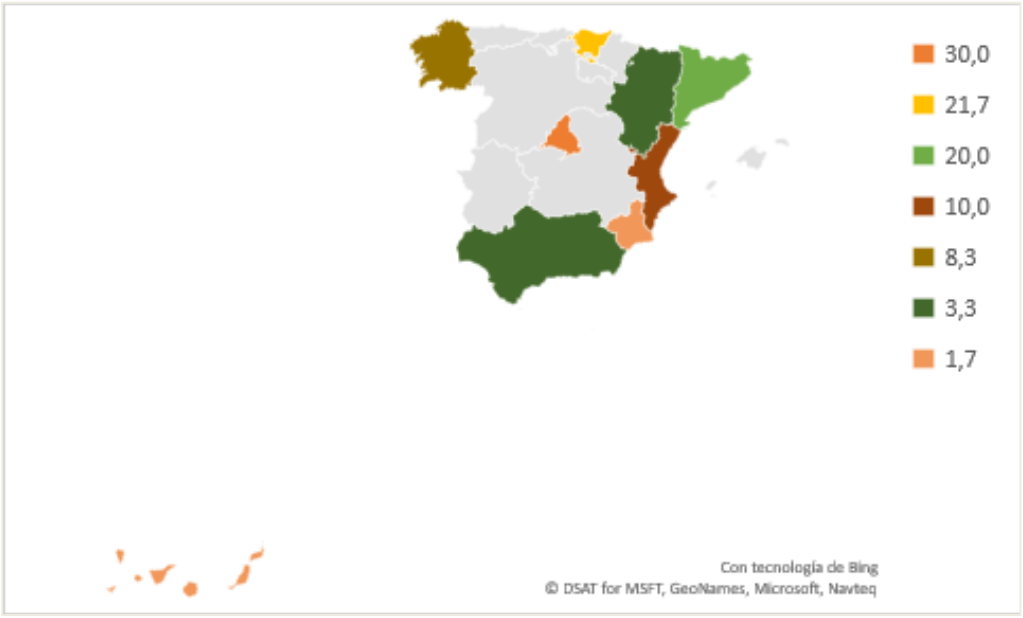
FIGURA 3. ÁMBITO GEOGRÁFICO AGENTES DEL SECTOR %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. F7

A continuación, se muestra un mapa de distribución geográfica de los agentes identificados del sector:

FIGURA 4. MAPA GEOGRÁFICO AGENTES DEL SECTOR %



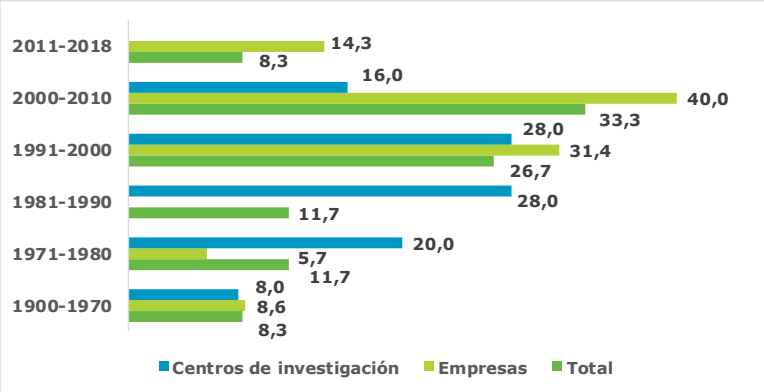
Antigüedad de los agentes del sector

En lo que respecta a la antigüedad de los agentes del sector se podría afirmar que se trata de empresas y centros de investigación relativamente jóvenes, ya que el 33,3% de los agentes iniciaron su actividad entre 2000 y 2010, y el 26,7% lo hicieron entre 1991 y 2000.

En el caso de las empresas, el 40% inició su actividad entre 2000 y 2010 y el 31,4% entre 1991 y 2000. **En comparación con los centros de investigación parece tratarse de empresas más jóvenes**, ya que el 28% de los centros inició su actividad entre 1991 y 2000, y el 28% lo hizo entre 1981 y 1990.

Además, hasta un 20% de los centros de investigación consultados iniciaron su actividad entre 1971 y 1980 mientras tan solo el 5,7% de las empresas iniciaron su actividad en esa década.

FIGURA 5. AÑO INICIO DE LA ACTIVIDAD AGENTES DEL SECTOR %

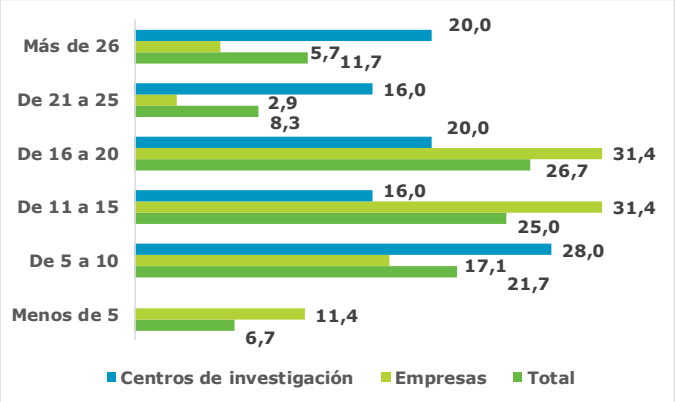


Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. F5T2

En lo que respecta a la antigüedad de la actividad relacionada con las tecnologías del lenguaje, el 51,7% de los agentes del sector lleva entre 11 y 20 años dedicándose a la actividad del sector de tecnologías del lenguaje, lo que implica cierta madurez de la actividad.

Si comparamos las empresas y los centros de investigación del sector podemos comprobar como las empresas llevan dedicándose menos tiempo a actividades del sector que los centros, un 11,4% lleva menos de 5 años dedicándose a actividades del sector de tecnologías del lenguaje frente a ningún centro de investigación. Por otra parte, tan solo el 8,6% de las empresas del sector lleva más de 21 años dedicándose a este tipo de actividad frente a un 36% de los centros de investigación, lo que indica que los centros de investigación son más antiguos que las empresas y que, además, llevan más tiempo dedicándose al sector de tecnologías del lenguaje.

FIGURA 6. ANTIGÜEDAD DE LA ACTIVIDAD DE TECNOLOGÍAS DEL LENGUAJE (AÑOS) %



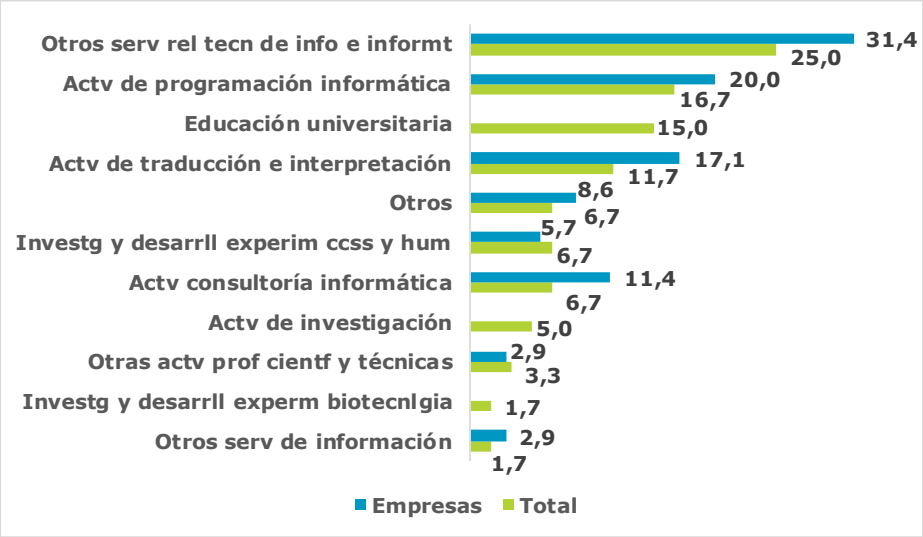
Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. F6

Distribución sectorial de la actividad

Los agentes del sector consultados expresaron que la actividad a la que se dedicaban se podía clasificar hasta en 11 CNAE distintos. El 25% de los agentes consultados expresaron que su actividad se correspondía con el CNAE 6209 "Otros servicios relacionados con las tecnologías de la información y la informática", el 16,7% afirmaron dedicarse a "Actividades de programación informática" (CNAE 6201) y el 15% a "Educación universitaria" (CNAE 8543).

Cabe señalar, que la mayoría de las empresas consultadas se dedican a "Otros servicios relacionados con las tecnologías de la información y la informática (31,4%), en segundo lugar, destacan las "Actividades de programación informática" y en tercer lugar las "Actividades de traducción e interpretación" (17,1%) (CNAE 7430).

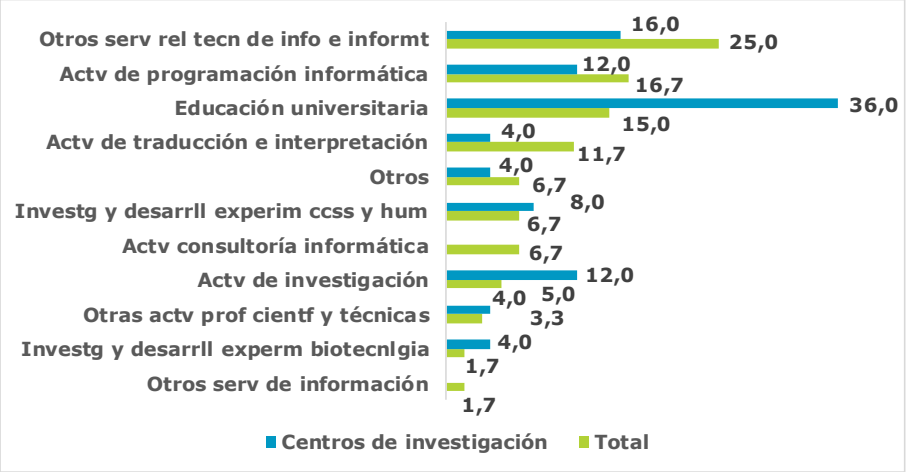
FIGURA 7. ACTIVIDAD DE LAS EMPRESAS DEL SECTOR %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35. F4

Mientras, los centros de investigación señalaron en su mayoría dedicarse a la "Educación universitaria" (36%), en segundo lugar y tercer lugar coinciden con las empresas en señalar las "Otros servicios relacionados con las tecnologías de la información y la informática" (16%) y "Actividades de programación informática" (12%).

FIGURA 8. ACTIVIDAD DE LOS CENTROS DE INVESTIGACIÓN DEL SECTOR %

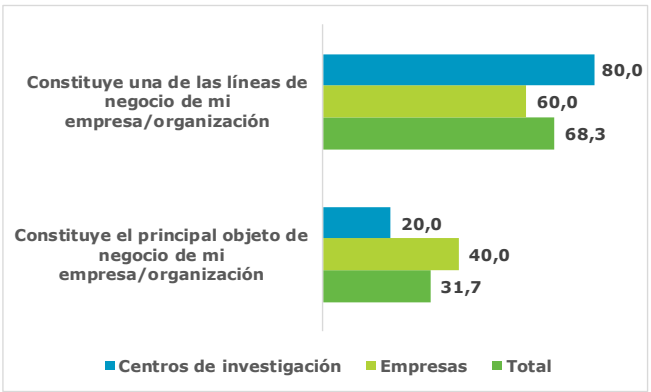


Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Centros de investigación 25. F4

El sector de tecnologías del lenguaje como actividad principal

La mayoría de los agentes del sector consultados se dedican a la actividad de tecnologías del lenguaje como una de las líneas de negocio de su empresa u organización, el 68,3%, lo que podría mostrar que es un tipo de actividad que se combina con otro tipo de actividades TIC, como muestran los sectores de actividad a los que se dedican empresas y centros de investigación. Llama la atención que la gran mayoría de los centros de investigación consultados, el 80%, indicaron que las tecnologías del lenguaje constituyen una de las líneas de negocio o investigación de su organización, lo que se podría explicar por la carga docente vinculada a la Universidad, por la que los centros de investigación combinan educación con investigación.

FIGURA 9. ACTIVIDAD DEL SECTOR COMO PRINCIPAL OBJETO DE LOS AGENTES DEL SECTOR %



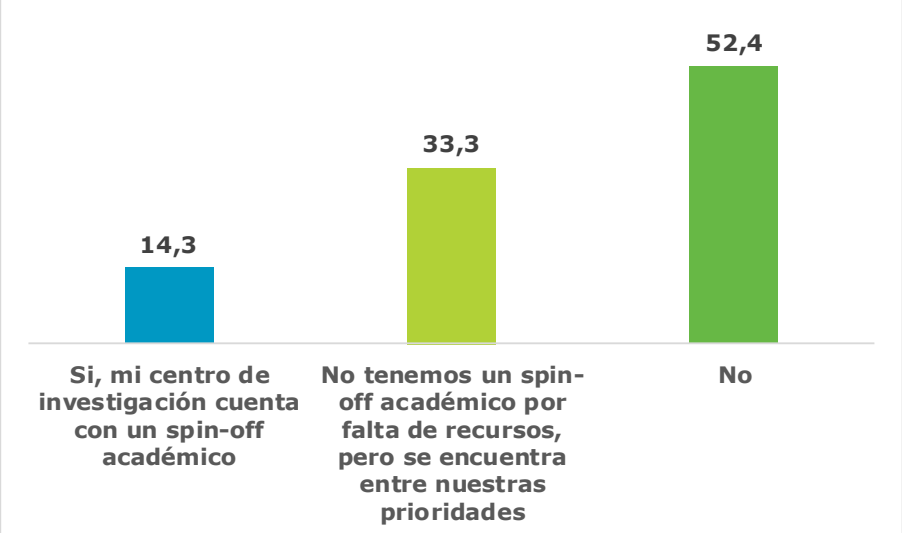
Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P1

Cabe destacar que el 66% de las empresas entre 50 y 249 empleados se dedica exclusivamente a actividades del sector de tecnologías del lenguaje. Por otra parte, el 75% de las empresas que se dedican a actividades de programación informática expresaron que la actividad de tecnologías del lenguaje es el principal objeto de su negocio. En este sentido, entre los agentes consultados las empresas medianas y que se dedican a actividades de programación informática se dedican en mayor proporción a la actividad TL como principal objeto de su negocio.

Para terminar, se preguntó a los centros de investigación si habían creado una spin-off para comercializar algún producto o servicio que hubieran desarrollado. Tan solo el 14,2% de los centros de investigación consultados ha creado una spin-off, lo que indica que los centros están orientados fundamentalmente a la investigación.

Además, el 33% de los centros de investigación afirmaron no tener un spin-off académico por falta de recursos, pero expresaron que se encuentra entre sus prioridades, lo que contrasta con la información obtenida en las entrevistas en profundidad que se realizaron a las administraciones, donde indicaban que los centros de investigación tienen complicaciones para crear spin-off vinculadas con la necesidad de invertir gran cantidad de tiempo, que han de dedicar a la investigación, por lo que no encontrarían incentivos para crear una spin-off.

FIGURA 10. CENTROS DE INVESTIGACIÓN QUE HAN CREADO SPIN-OFF %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Centros de investigación 21. P13

El censo identificado del sector está conformado por 128 empresas y 63 centros de investigación.

En general, el régimen jurídico de las empresas es sociedad de responsabilidad limitada y los centros de investigación forman consorcios.

Los agentes del sector se distribuyen mayoritariamente entre la Comunidad de Madrid, el País Vasco, Cataluña y Comunidad Valenciana.

Los agentes del sector tienen cierta madurez en la dedicación a actividades de tecnologías del lenguaje, aunque la mayoría desarrolla la actividad TL combinada con otras líneas de negocio.

Los centros de investigación están orientados fundamentalmente a la investigación y en su mayoría no llegan a crear *spin-offs* para comercializar componentes de productos o aplicaciones.

La mayoría de los agentes consultados dedica su actividad a "Otros servicios relacionados con las tecnologías de la información y la informática", "Actividades de programación informática" y "Educación universitaria".

2.2 Personal ocupado del sector

Tamaño de los agentes del sector

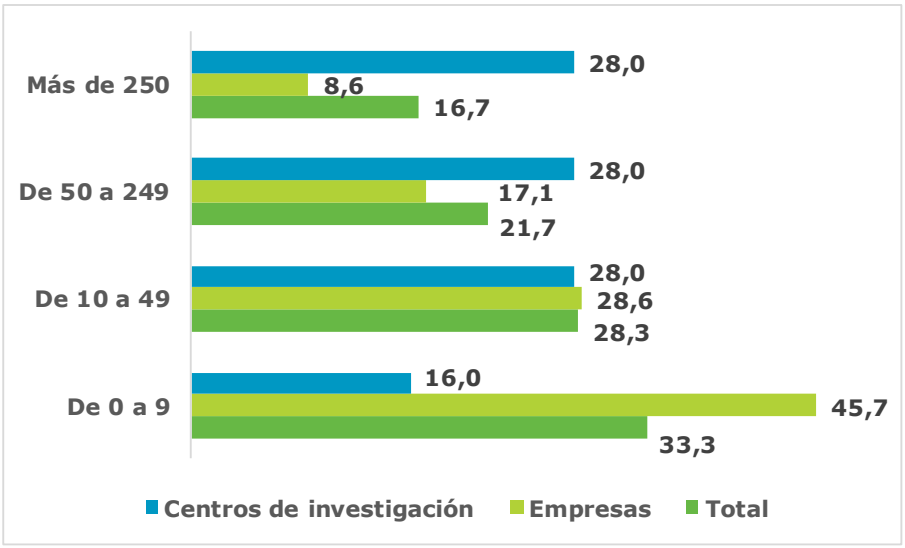
Poco más del **60% del total de los agentes del sector consultados tiene menos de 49 empleados**. El porcentaje aumenta hasta el 74,3% si hablamos de empresas, mientras los centros de investigación con menos de 49 empleados no llegan al 50%.

Esto viene motivado porque algunos centros de investigación pertenecen a universidades, por lo que el número de empleados de su organización está vinculado a la Universidad a la que pertenecen. En este sentido, no resulta representativo hablar de un 28% de centros de investigación con más de 250 empleados, si no se entiende que se refiere a la Universidad en su conjunto.

Más aún si se tiene en cuenta la información aportada por los centros de investigación en los grupos de discusión donde afirmaron que *"un porcentaje muy alto de la investigación en la Universidad se hace en estructuras muy pequeñas".*

De cualquier modo, **se trata de un sector de actividad concentrado en microempresas y pymes**, tan solo el 17,1% de las empresas consultadas son medianas y alrededor del 8% son grandes empresas.

FIGURA 11. NÚMERO DE EMPLEADOS DE LOS AGENTES DEL SECTOR %

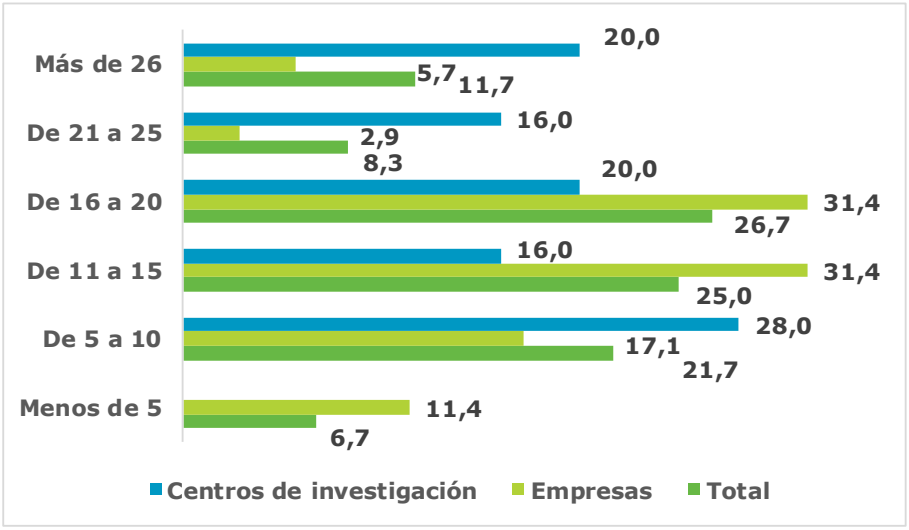


Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P2A

Volumen de empleados vinculados a la actividad de TL

El 73% de los agentes consultados manifestó que el volumen de empleados de su empresa u organización vinculados a la actividad de tecnologías del lenguaje se encuentra **entre los 5 y los 20 empleados.**

FIGURA 12. NÚMERO DE EMPLEADOS RELACIONADOS CON LA ACTIVIDAD DEL SECTOR DE TECNOLOGÍAS DEL LENGUAJE %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Centros de investigación 21. P2B

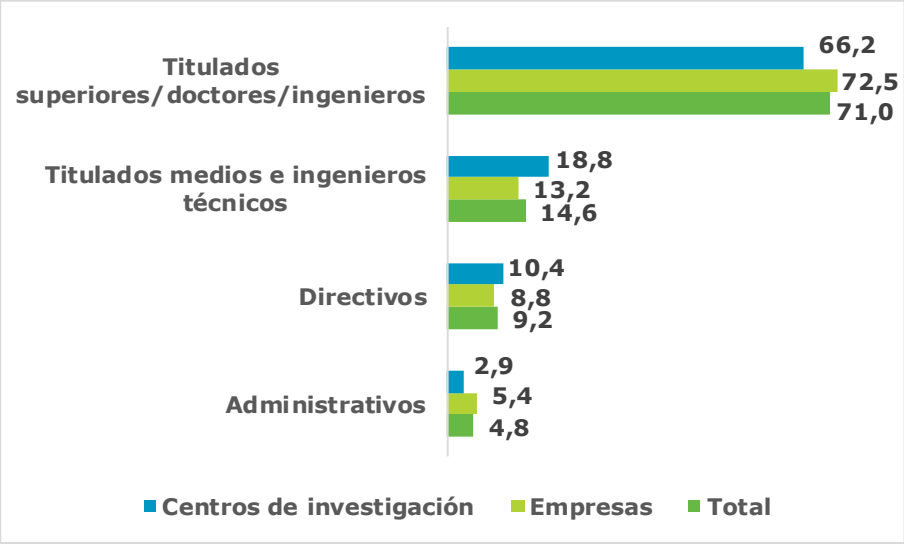
El 62,9% de las empresas consultadas afirmó que el grueso de empleados vinculados a la actividad del sector de tecnologías del lenguaje se encuentra entre 11 y 20 empleados. Por otra parte, destaca que el 20% de los centros de investigación expresó que tienen más de 26 empleados vinculados con las actividades de tecnologías del lenguaje, con lo que parecería que están más orientados a este tipo de actividades.

Además, en los grupos de discusión, los centros de investigación manifestaron que sus plantillas están formadas por personal permanente y por una gran cantidad de becarios, que podrían contribuir a aumentar las cifras de empleados vinculados con el sector: *"Al ser un grupo de investigación esto fluctúa, está el personal permanente, y está el personal de becarios que son los que realmente hacen el trabajo, es la gente que mantiene la investigación en este país".*

Categoría profesional de los empleados del sector

El 70,9% de los empleados del sector de los agentes consultados son **titulados superiores, doctores o ingenieros**. La distribución es muy similar para empresas y centros de investigación, lo que indica que las actividades del sector de tecnologías del lenguaje están muy vinculadas a formación superior.

FIGURA 13. EMPLEADOS ASOCIADOS A ACTIVIDADES RELACIONADAS CON LAS TECNOLOGÍAS DEL LENGUAJE POR CATEGORÍA PROFESIONAL %

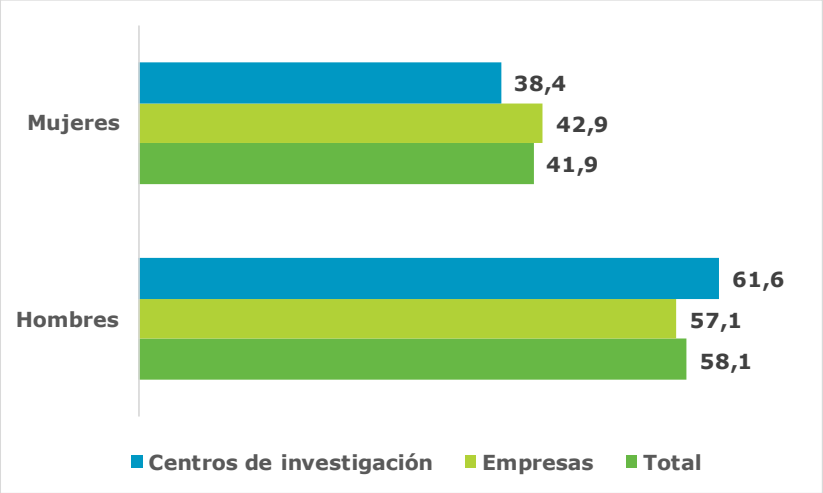


Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P5

Categoría profesional desde la perspectiva de género

Por otra parte, un 58,1% de los empleados de los agentes del sector son hombres, siendo este porcentaje algo más alto en los centros de investigación (61,6%) que en las empresas (57,1%).

FIGURA 14. EMPLEADOS ASOCIADOS A ACTIVIDADES RELACIONADAS CON LAS TECNOLOGÍAS DEL LENGUAJE POR GÉNERO %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P6

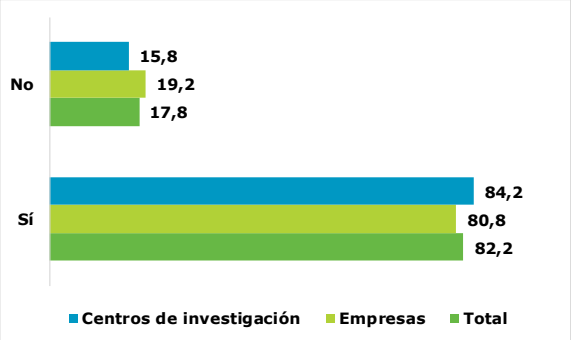
En las entrevistas en profundidad realizadas a empresas y centros de investigación se les preguntó acerca de su opinión respecto a la posible existencia de la denominada brecha de género en el sector.

La mayoría de las empresas y centros de investigación afirmó que perciben una brecha de género menor a la existente en otros sectores debido a que la actividad de tecnologías del lenguaje es multidisciplinar, en la medida que precisa de lingüistas y de programadores o técnicos.

Contratación de personal en el sector

El 75% de los agentes consultados afirmó haber contratado personal durante el año 2017, cifra muy similar entre empresas (74,2%) y centros de investigación (76%). Estos datos **podrían indicar un crecimiento de la demanda de la industria de tecnologías del lenguaje**. Además, el 80,8% de las empresas y el 84,2% de los centros de investigación consultados que contrató personal en el año 2017 manifestaron que era personal especializado en actividades relacionadas con el procesamiento del lenguaje natural, la traducción automática y los sistemas conversacionales.

FIGURA 15. ¿HA CONTRATADO PERSONAL ESPECIALIZADO EN ACTIVIDADES RELACIONADAS CON LAS TL? %

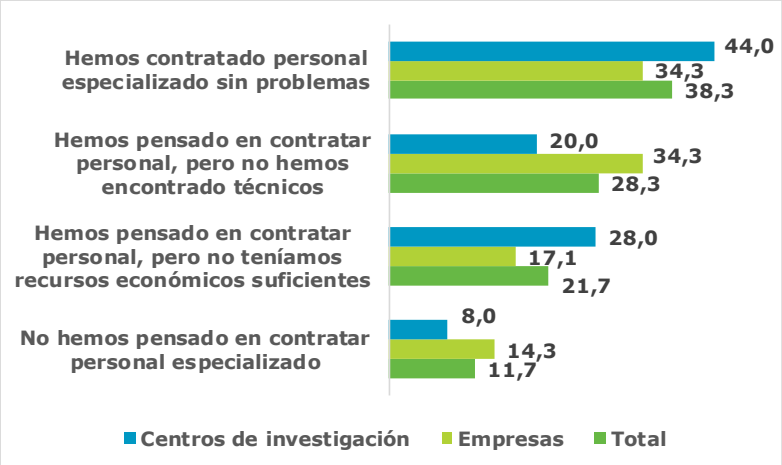


Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 45, Empresas 26, Centros de investigación 19. P3A

No obstante, se preguntó a los agentes del sector por las posibles barreras que hubieran tenido a la contratación, independientemente de si habían contratado personal especializado en el sector de tecnologías del lenguaje. A este respecto, un 28,3% de los agentes expresaron haber pensado en contratar personal y haber tenido dificultad para encontrar los técnicos adecuados y un 21,6% aludió no haber tenido los recursos económicos suficientes.

Parece que las empresas han tenido más problemas para encontrar técnicos (un 34,2%) que los centros de investigación (20%), mientras estos últimos habrían tenido más problemas económicos (28%) que las empresas (17,1%).

FIGURA 16 ¿EN ALGÚN MOMENTO HAN PENSADO EN CONTRATAR PERSONAL ESPECIALIZADO? %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P4

Con el fin de matizar las respuestas del cuestionario web se preguntó a las empresas en la entrevista en profundidad y en el grupo de discusión si la demanda de profesionales en el sector de las tecnologías del lenguaje está cubierta en relación con la oferta del mercado.

La mayoría de las empresas expresó que la demanda de profesionales en el sector de tecnologías del lenguaje no está cubierta por la oferta del mercado fundamentalmente por un problema de formación. La especificidad del sector reside en la necesidad de encontrar perfiles mixtos con conocimientos en tecnología y conocimientos en lenguaje e idiomas.

En este sentido, se detectó una falta de combinación de la parte de conocimiento lingüístico con la parte técnica que no se encuentra integrada en las titulaciones técnicas, por lo que, a menudo, son las empresas las que asumen la formación de sus empleados técnicos en el área lingüística, o a los lingüistas en el área técnica:

"Especialmente en España falta personal cualificado, la demanda es mayor que la oferta, dado que las empresas no encuentran personal cualificado han de formar a los empleados que contratan en base a los perfiles específicos que precisa su actividad".

"En nuestra empresa ha sido un reto, conviven un técnico, una lingüista, un técnico, una lingüista, como quieran, pero colaboran, y esa es una labor que estamos haciendo las empresas. Yo vengo del área lingüística y en varios procesos tengo que hacer de intermediario o de traductor, entre el área de ingeniería y el área de programación, y el área de la lingüística, porque no existe formación en esta materia. Los ingenieros saben de programación, los lingüistas saben de semántica, de fonética, de crear diálogos, pero luego no saben cómo unir las piezas".

Además, un grupo de empresas señalaron que el problema está en la formación de los técnicos en el área de lingüística más que en la formación de los lingüistas en el área técnica:

"El problema está fundamentalmente en la formación de los técnicos, no de los lingüistas, esto ha podido verse influenciado por la limitada salida académica que tenían las carreras relacionadas con la lingüística, que ha llevado a los lingüistas a investigar otras salidas profesionales, otros enfoques a sus conocimientos. Sin embargo, los ingenieros y programadores han sido profesiones más demandadas por el mercado y no han tenido esa necesidad de formarse en otros campos como es la lingüística".

Algunas empresas argumentaron que, al tratarse de un mercado pequeño que se está renovando continuamente, están surgiendo nuevas tecnologías a menudo, por lo que el sector precisa de profesionales formados que sean capaces de desarrollar las diferentes aplicaciones tecnológicas.

Por otra parte, la opinión de los centros de investigación sirvió para contrastar lo expuesto por las empresas, ya que, en gran medida los centros de investigación están vinculados con la Universidad, por lo que tienen un papel importante en la formación de especialistas en tecnologías del lenguaje.

Por tanto, desde la perspectiva de la oferta de profesionales del sector, la mayoría de los centros de investigación entrevistados afirmó que las universidades no cuentan con titulaciones técnicas especializadas en tecnologías del lenguaje. Alguna universidad tiene un máster propio, pero solo con algunas asignaturas relacionadas con las tecnologías del lenguaje.

En este sentido, se recomienda la creación de una titulación técnica que integre los conocimientos técnicos y de programación, y los conocimientos lingüísticos necesarios para desarrollar la actividad de tecnologías del lenguaje.

Algunos centros de investigación argumentaron que el propio funcionamiento de la Universidad dificulta la adaptación a las necesidades de las empresas que, en sectores relacionados con este tipo de tecnologías tan avanzadas, están en continua renovación:

"Los mecanismos de formación que tiene la Universidad carecen de la agilidad que necesita esto, se tarda seis años en que se apruebe un plan de estudios y, conforme pasa el tiempo, en seis años las empresas tienen otras necesidades".

Por otra parte, desde la perspectiva de los centros de investigación como demandantes de especialistas, la mayoría señaló tener dificultades para contratar personal cualificado por problemas de financiación pública para contratar personal, ya que los perfiles que precisan son de cualificación muy alta y el sueldo que pueden ofrecer muy bajo en comparación con la empresa privada, por lo que los especialistas mejor formados aspiran a trabajar en la empresa antes que en el centro de investigación.

Esta falta de compensación entre centro de investigación y empresa también conlleva que sus esfuerzos de formación en especialistas no se vean recompensados ya que el contrato que ofrece la Universidad no es competitivo:

"Contratamos a estudiantes que han terminado el máster y les formamos nosotros mismos".

"Empiezas a formar a gente y cuando están formados, se te escapan a otro sitio".

La mayoría de las empresas consultadas tienen menos de 49 empleados, lo que mostraría un sector conformado por microempresas y pequeñas empresas.

Las cifras de empleados vinculados a la actividad de tecnologías del lenguaje muestran que las empresas que conforman el sector consultadas están orientadas parcialmente a la actividad de tecnologías del lenguaje, ya que el 62,9% afirmó que su plantilla de empleados vinculados a las tecnologías del lenguaje se encontraba entre 11 y 20 empleados.

Los agentes que conforman el sector consultados han expresado en su gran mayoría haber contratado personal especializado durante el último año (82,2%), lo que podría mostrar que la actividad de tecnologías del lenguaje está en auge, dado que las empresas y los centros de investigación están ampliando sus plantillas.

La actividad de tecnologías del lenguaje requiere formación universitaria superior, como muestra el 70,9% de los empleados vinculados con el sector con titulaciones superiores, doctorandos o ingenieros.

Por otra parte, aunque el sector está conformado por perfiles mixtos de técnicos y lingüistas, existe una brecha de género del 16,2%.

2.3 Oferta de soluciones de tecnologías del lenguaje

Calificación de productos y servicios del sector

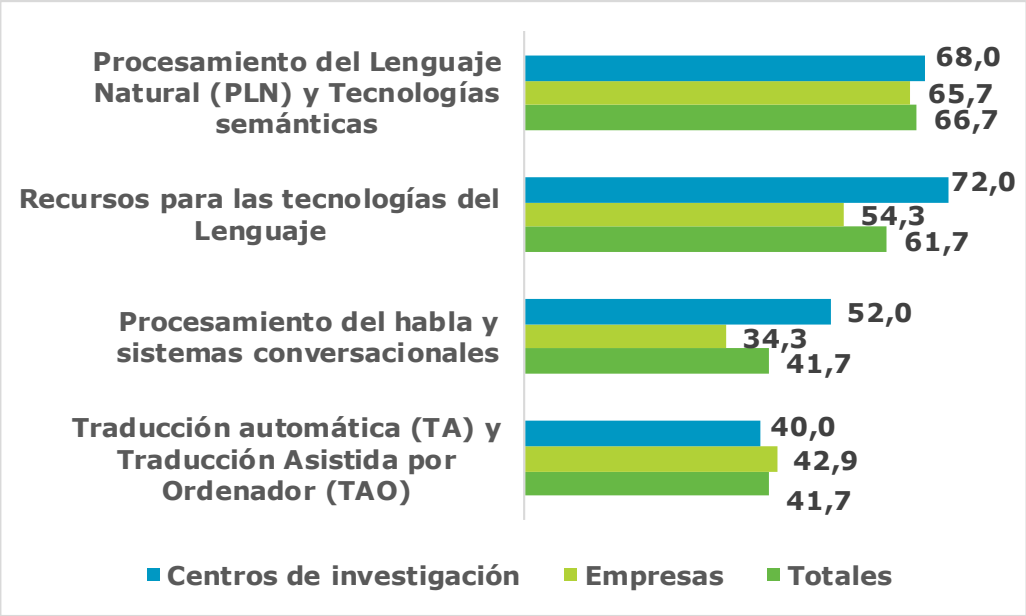
Las aplicaciones basadas en las tecnologías del lenguaje se clasifican en cuatro grandes tipos de actividades: la traducción automática y traducción asistida por ordenador, el procesamiento del habla y los sistemas conversacionales, el procesamiento del lenguaje natural y las tecnologías semánticas, y los recursos para las tecnologías del lenguaje.

De las cuatro categorías anteriores, los agentes consultados expresaron comercializar o desarrollar una media de 2 tipos de actividades, lo que indicaría que las actividades del sector de tecnologías del lenguaje están de alguna forma interconectadas y no existe una gran especialización por tipo de producto o servicio que se comercializa o desarrolla entre los agentes consultados.

La mayoría de los agentes del sector (66,7%) se dedican a actividades de procesamiento del lenguaje natural, en segundo lugar, el 61,7% se dedica a los recursos para las tecnologías del lenguaje, lo que guarda cierta lógica ya que son la base del desarrollo de este tipo de tecnologías. En tercer y cuarto lugar encontraríamos las actividades de traducción automática y traducción asistida por ordenador con un 41,7%.

Si comparamos lo expresado por empresas y centros de investigación se puede señalar que los centros de investigación se dedican en mayor medida a actividades de procesamiento del habla y sistemas conversacionales (un 52% frente a un 34,3% de las empresas). También lo hacen en mayor medida en el caso de los recursos para las tecnologías del lenguaje (72% frente a un 54,3% de las empresas).

FIGURA 17. TIPOLOGÍA DE PRODUCTOS QUE COMERCIALIZAN LOS AGENTES %

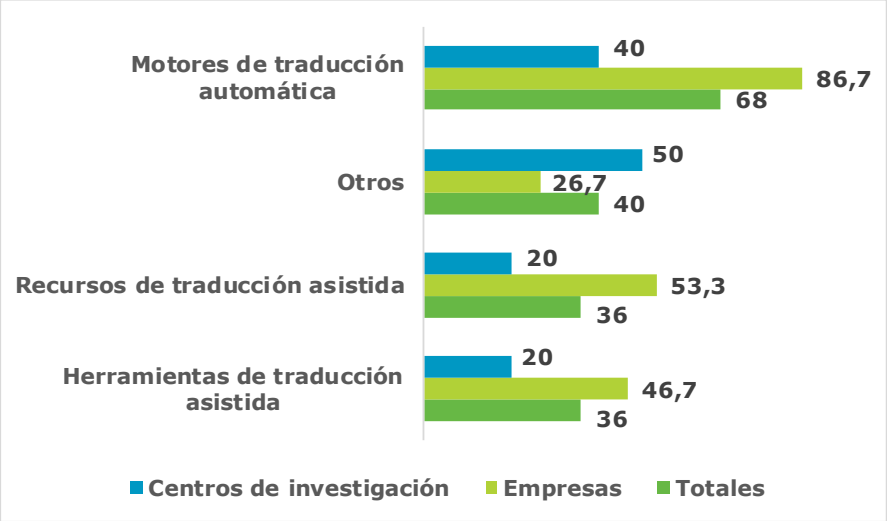


Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P7

A continuación, se presenta la distribución de las respuestas que han dado los agentes del sector sobre las herramientas de tecnologías del lenguaje vinculadas a cada una de las categorías de producto precedentes.

En el caso de las soluciones de traducción automática y traducción asistida por ordenador las herramientas sobre las que se ha preguntado son: motores de traducción automática, recursos de traducción asistida y herramientas de traducción asistida.

FIGURA 18. TIPOLOGÍA DE HERRAMIENTAS DE TRADUCCIÓN AUTOMÁTICA QUE COMERCIALIZAN LOS AGENTES %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P7A

El 68% de los agentes que comercializan o investigan productos o servicios de traducción automática y traducción asistida por ordenador, **desarrollan motores de traducción automática**. En el caso de las empresas consultadas la especialización en este tipo de productos o servicios es todavía mayor (un 86,7%).

Cabe señalar que un 50% de los centros de investigación expresaron comercializar con otro tipo de productos o servicios, entre los que se encuentra la clasificación automática de textos, la recuperación de información, corpus anotados automáticamente, POS targers, parsers[2] y actividades de evaluación de la calidad de la voz para usos clínicos.

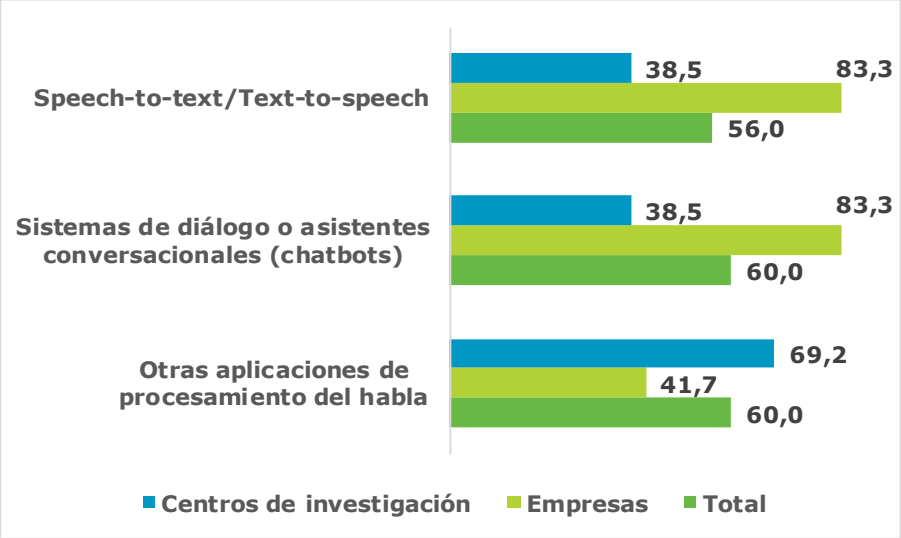
En el caso de las soluciones relacionadas con el procesamiento del habla y los sistemas conversacionales las herramientas por las que se ha preguntado son: sistemas de diálogo o asistentes conversacionales (chatbots) y speech-to-text/text-to-speech.

[2] Part off speach targger. Son trozos de software que leen en un determinado idioma y asignan a cada palabra su función como parte d discurso (nombre, verbo, adjetivo, etc.). Un analizador sintáctico (o parser) es un programa informático que analiza una cadena de símbolo de acuerdo a las reglas de una gramática formal.

La mayoría de los agentes consultados desarrollan las dos herramientas básicas de procesamiento del habla y sistemas conversacionales.

El 60% desarrolla sistemas de diálogo o asistentes conversacionales y speech-to-text/text-to-speech, en último lugar, el 56% desarrolla otras aplicaciones de procesamiento del habla.

FIGURA 19. TIPOLOGÍA DE HERRAMIENTAS DE SISTEMAS CONVERSACIONALES QUE COMERCIALIZAN LOS AGENTES %

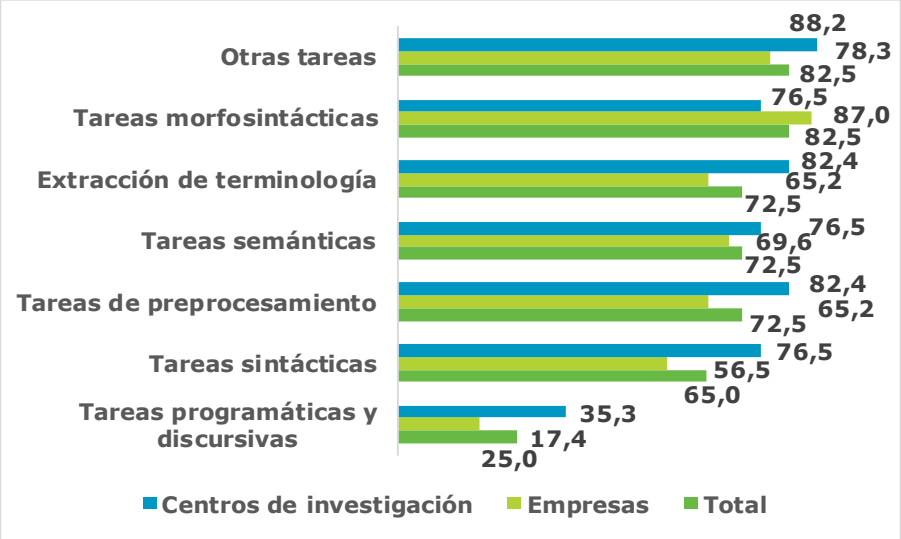


Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P7B

Las otras aplicaciones de procesamiento del habla mencionadas por las empresas y los centros de investigación son los sistemas de biometría vocal y *speech analytics* y sistemas de detección de emociones y análisis del sentimiento. Por otra parte, las empresas y los centros de investigación han expresado desarrollar aplicaciones de reconocimiento del hablante y la lengua.

En el caso de las soluciones de procesamiento del lenguaje natural, la mayoría de los agentes consultados comercializa o desarrolla tareas de preprocesamiento, tareas morfosintácticas, tareas sintácticas, tareas semánticas, extracción de terminología u otras tareas (reconocimiento de entidades nombradas y su clasificación, entity linking, extracción de relaciones, asistencia a la redacción, sistemas de asistencia a la pronunciación, polaridad, etc.).

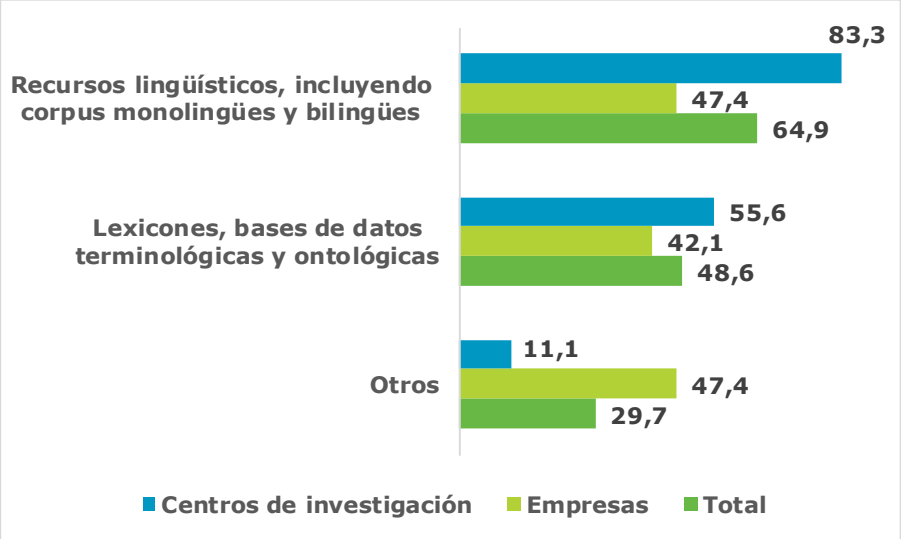
FIGURA 20. TIPOLOGÍA DE TAREAS DE PROCESAMIENTO DE LENGUAJE NATURAL QUE COMERCIALIZAN LOS AGENTES %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P7C

La actividad que menos desarrollan los agentes del sector es la relacionada con tareas programáticas y discursivas con un 25%. Este tipo de tareas se da en mayor proporción en los centros de investigación (35,3%) que en las empresas (17,4%).

FIGURA 21. TIPOLOGÍA DE HERRAMIENTAS DE RECURSOS LINGÜÌSTICOS QUE COMERCIALIZAN LOS AGENTES%



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P7D

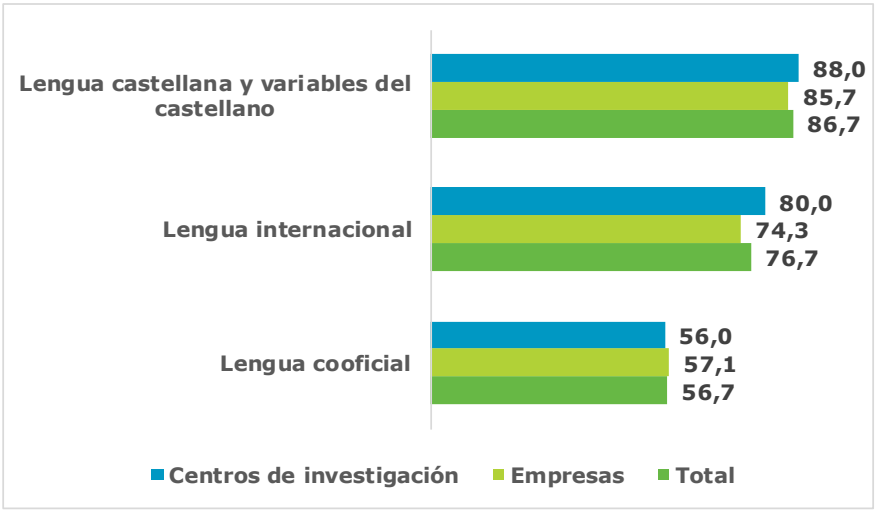
Para terminar, el 64,9% de los agentes consultados que comercializan recursos para las tecnologías del lenguaje, producen recursos lingüísticos, incluyendo corpus monolingües y bilingües. Mientras, el 48,6% de los agentes produce lexicones y bases de datos terminológicas y ontológicas.

Los agentes que han expresado producir o comercializar otro tipo de recursos lingüísticos han especificado el uso que les dan a esos recursos de tecnologías del lenguaje. Los usos más mencionados por los agentes son: una plataforma cognitiva para sistemas conversacionales, software basado en tecnologías de procesamiento del lenguaje natural, traducción e interpretación, software de búsqueda semántica o documentación técnica.

Lenguas en las que se desarrolla la actividad

Respecto a las lenguas en las que orientan la actividad los agentes del sector, la mayoría de las empresas y los centros de investigación consultados (86,7%) afirmaron orientar su actividad a la lengua castellana y variables del castellano (como el colombiano o el venezolano). En un porcentaje menor (76,7%) los agentes del sector manifestaron orientar su actividad hacia alguna lengua internacional y, por último, algo más de la mitad orienta su actividad hacia alguna lengua cooficial. El sector de tecnologías del lenguaje es multilingüe, dado que la mayoría de los agentes consultados dirige su actividad en alguna lengua internacional.

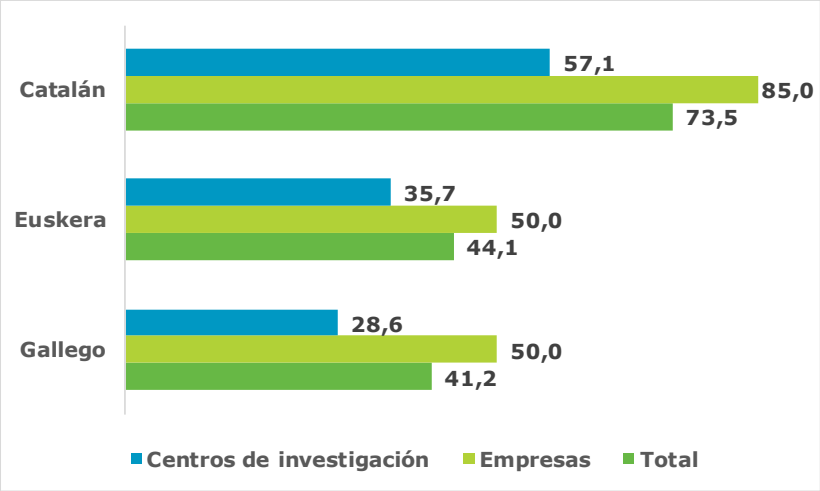
FIGURA 22. LENGUAS EN LAS QUE DESARROLLAN LA ACTIVIDAD LOS AGENTES %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P8

La lengua cooficial hacia la que más dirigen su actividad los agentes identificados es el catalán en primer lugar con un 73,5%, mientras el euskera y gallego es utilizado por alrededor del 40% de los agentes del sector consultados.

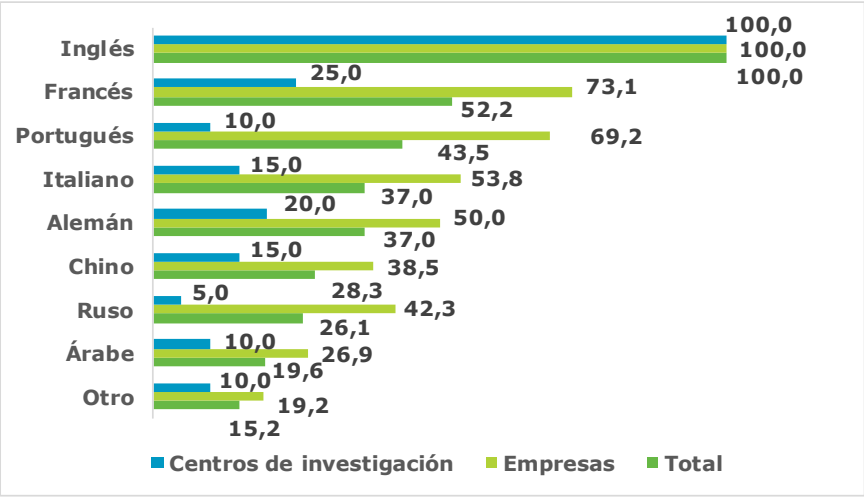
FIGURA 23. LENGUAS NACIONALES EN LAS QUE DESARROLLAN LA ACTIVIDAD LOS AGENTES %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P8A

En lo que respecta a las lenguas internacionales, la mayoría de los agentes consultados expresó orientar la actividad de su negocio o investigación en, al menos, tres lenguas distintas.

FIGURA 24. LENGUAS INTERNACIONALES EN LAS QUE ORIENTAN LA ACTIVIDAD LOS AGENTES %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 60, Empresas 35, Centros de investigación 25. P8B

Todos los agentes que dirigen su actividad a lenguas internacionales manifestaron hacerlo en la lengua inglesa, lo que muestra el gran predominio de este idioma en el mercado. La segunda lengua hacia la que dirigen más su actividad los agentes del sector consultados es el francés con un 52,2% y en tercer lugar el portugués con un 43,5%. En cuarto lugar, encontraríamos el italiano y el alemán con un 37%.

Cabe destacar que después del inglés, las empresas orientan su actividad en mayor proporción hacia el francés (73,1%), el portugués (69,2%) y el italiano (53,8%), lo que indicaría la importancia de la proximidad de las lenguas para la orientación comercial de las empresas del sector. Por su parte, los centros de investigación dirigen de forma muy residual sus investigaciones hacia otros idiomas que no sean el inglés.

En las entrevistas y los grupos de discusión se preguntó a las empresas y los centros de investigación por la situación de las lenguas cooficiales respecto al castellano y, por lo general, expresaron que en comparación con la lengua castellana faltan recursos para el catalán, euskera o gallego, es decir, faltan corpus que sean representativos de estos idiomas.

No obstante, las herramientas y los recursos de las distintas lenguas cooficiales son considerables y, el hecho de que España sea un país multilingüe ha permitido que el sector de tecnologías del lenguaje avance.

La mayoría de las empresas y centros de investigación coincidieron en señalar al catalán como la lengua cooficial que más recursos y herramientas tiene, seguida del euskera y, en último lugar, el gallego. En este sentido, afirmaron que el impulso de las administraciones autonómicas en las tecnologías del lenguaje es esencial en el desarrollo de las lenguas cooficiales.

Algunas empresas señalaron que en Europa empieza a haber cierto interés en las lenguas cooficiales que se hablan en los distintos estados miembros, en tanto una parte del mercado y la población está orientado en esas lenguas:

"En Europa se empieza a hablar de las otras lenguas oficiales. En algunos estados miembros se empieza a valorar la importancia de lenguas que no solamente se hablan en los hogares, sino que son lenguas que tienen su lugar en la enseñanza y se usan en ámbitos comerciales".

Por otra parte, se preguntó a las empresas y los centros de investigación por la situación del castellano a nivel europeo. La mayoría de las empresas y centros de investigación consultados coincidió en posicionar a la lengua castellana a niveles similares que otras lenguas hegemónicas, como el inglés, el

francés y el alemán, ya que consideran que existen bastantes recursos y está bien caracterizada porque existen bastantes herramientas y tiene presencia en el sector.

No obstante, la lengua inglesa es la predominante, desde el punto de vista de desarrollo de soluciones, de investigación y de comercialización.

"Existen bastantes recursos, bases de datos de español, lo que pasa es que los recursos que hay disponibles de inglés son muy superiores, bastante más que la proporción que existe de número de hablantes".

Algunas empresas expresaron que la lengua castellana podría avanzar más en el sector de tecnologías del lenguaje en el camino a conseguir unas infraestructuras lingüísticas genéricas:

"Las tecnologías del lenguaje son soluciones que requieren muchísimos recursos, mucho conocimiento, mucha base de datos, mucha gramática, en ese aspecto, aunque se ha hecho mucho esfuerzo por generar herramientas libres, mi percepción es que no estamos todavía en condiciones de hacer un trabajo con recursos que sean genéricos".

Las empresas y centros de investigación señalaron que la situación de la lengua castellana a nivel internacional mejora respecto a su situación a nivel europeo gracias a la introducción del mercado latinoamericano, que la mayoría coincide en señalar como un nicho de mercado del sector. No obstante, señalaron la necesidad de diferenciar el castellano del español de Latinoamérica, especialmente en el procesamiento del habla y en los sistemas conversacionales.

"En el caso de las tecnologías del habla para el reconocimiento de voces y para la conversión de voz a texto, España está bastante cubierta, pero en el momento que hacemos el salto a Latinoamérica queda muchísimo trabajo por hacer, porque se ha considerado que español es todo y las formas de realización en los distintos países de Latinoamérica son muy diferentes. Es una asignatura que tenemos pendiente".

Por otra parte, algunas empresas y centros de investigación señalaron que el mercado norteamericano es el que está copando el mercado internacional del sector, tanto para el inglés como para el castellano, debido a la gran presencia en el sector de sus multinacionales.

"El problema está en que la lengua castellana, aunque tenga una amplia penetración en distintos países no tiene un predominio tecnológico en lo que respecta a las soluciones del lenguaje natural, son los proveedores de tecnologías norteamericanos los que están proporcionando las mejores soluciones de español desde una perspectiva global. Pongo ejemplos, ahora mismo el mejor procesador de lenguaje

natural que existe para su integración y por su funcionamiento pertenece a empresas como Microsoft o Google".

"Comparado con el inglés todo es poco, el inglés avanza mucho más que el resto de los idiomas, y luego hay como diez o doce idiomas que estarían en primera fila detrás del inglés, que son los idiomas que utilizan las grandes empresas y que tienen más usuarios en Internet. El español está entre ellos y tenemos que hacer frente a las multinacionales".

Por tanto, **a** nivel internacional parece tratarse de un problema de tamaño de mercado, de grandes empresas norteamericanas frente a pequeñas empresas españolas, más que un problema de predomino lingüístico del inglés frente al español. El hecho de que las grandes empresas norteamericanas tengan más capacidad para desarrollar soluciones de tecnologías del lenguaje porque cuentan con mayor cantidad de corpus, apoyadas por una legislación en materia de protección de datos más flexible, justifica la necesidad de generar corpus en castellano a nivel nacional de manera que las empresas españolas puedan beneficiarse de estos recursos, tanto en castellano, como en las distintas lenguas cooficiales, para entrenar sus sistemas y aumentar su competitividad frente a las grandes empresas tecnológicas americanas en el mercado del español.

La mayoría de los agentes del sector comercializan con dos o más tipos de productos y servicios de tecnologías del lenguaje, lo que indicaría que este tipo de actividades están interconectadas e intervienen de manera horizontal en el desarrollo de soluciones.

El sector de las tecnologías del lenguaje es multilingüe, la mayoría de los agentes consultados orientan la actividad de su negocio o investigación en, al menos, tres lenguas distintas.

Las empresas se orientan en mayor proporción que los centros de investigación hacia lenguas de países próximos.

Por otra parte, la lengua cooficial hacia la que más dirigen su actividad los agentes consultados es el catalán.

2.4 Volumen de ventas del sector

Volumen de facturación del sector

Para el cálculo del volumen de facturación del sector se obtuvieron los datos a partir del Registro Mercantil para los años 2014 y 2015. Para calcular el volumen de facturación del año 2016 se estimó a partir de la evolución de los dos años anteriores, completando la información con los datos obtenidos

a partir de la encuesta en lo que se refiere al porcentaje que representan los ingresos derivados de las tecnologías del lenguaje.

El volumen de facturación que se corresponde con las actividades de tecnologías del lenguaje se situó alrededor de los 205 millones de euros.

Ingresos asociados a compras públicas

En las entrevistas en profundidad realizadas a las empresas, la mayoría afirmó no realizar ventas al sector público:

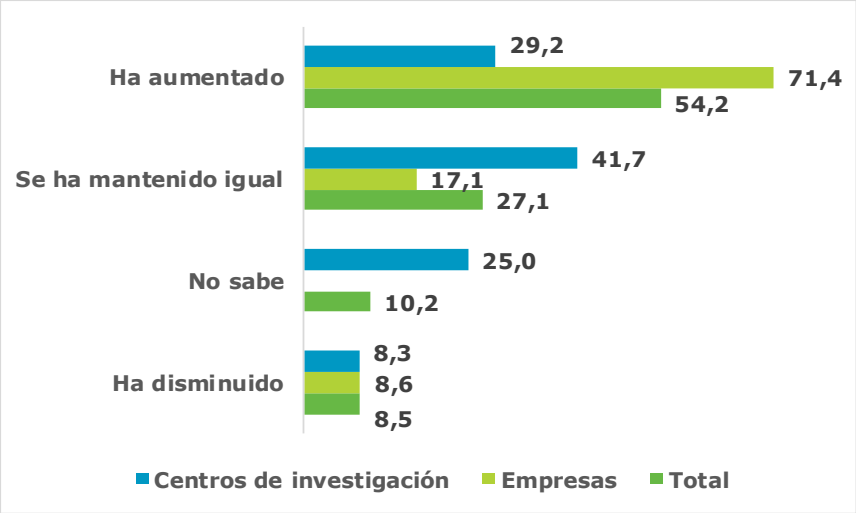
"Siempre vamos a un negocio más de empresa a empresa, y de todas formas encontramos dificultades al ser una empresa pequeña, las compras públicas casi siempre se las llevan las empresas grandes que pueden hacer ofertas más a la baja, y luego acaban contratando o subcontratando empresas pequeñas para que hagan el trabajo".

Por el contrario, a lo expresado por las empresas, los centros de investigación realizan ventas al sector público a través de convenios con las administraciones, a nivel local, regional, nacional y europeo.

Evolución del volumen de clientes

Respecto a la evolución del volumen de clientes, un 54,2% de los agentes consultados expresó que durante el año 2017 el volumen de sus clientes aumentó, este porcentaje es bastante mayor entre empresas (71,4%) que entre centros de investigación (29,2%).

FIGURA 25. VOLUMEN DE CLIENTES QUE HA AUMENTADO EN 2017 %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 59, Empresas 34, Centros de investigación 25. P14

Estos datos refuerzan la idea de un sector de actividad en auge, dado que la gran mayoría de las empresas que lo conforman han aumentado su volumen de clientes durante 2017, lo cual es un dato positivo.

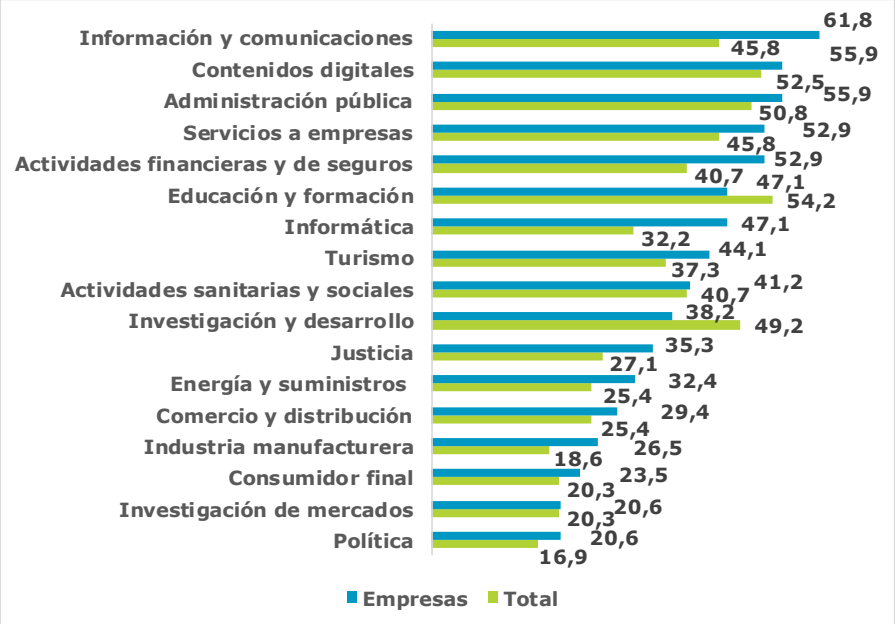
2.5 Destino funcional de las ventas del sector

Sectores demandantes de soluciones del sector

En términos generales los sectores que más destacan respecto a la demanda de soluciones de tecnologías del lenguaje son: el sector de Educación y formación, con un 54,2% de los agentes que dirigen sus ventas hacia este sector; el sector de contenidos digitales con un 52,2%, y el sector público con un 50,8%.

Las empresas consultadas dirigen sus ventas en mayor medida al sector de la Información y las comunicaciones con un 61,8%. También destacan el sector de Contenidos digitales y Administración pública con un 55,9% y, los sectores de Servicios a empresas y Actividades financieras y de seguros con un 52,9%.

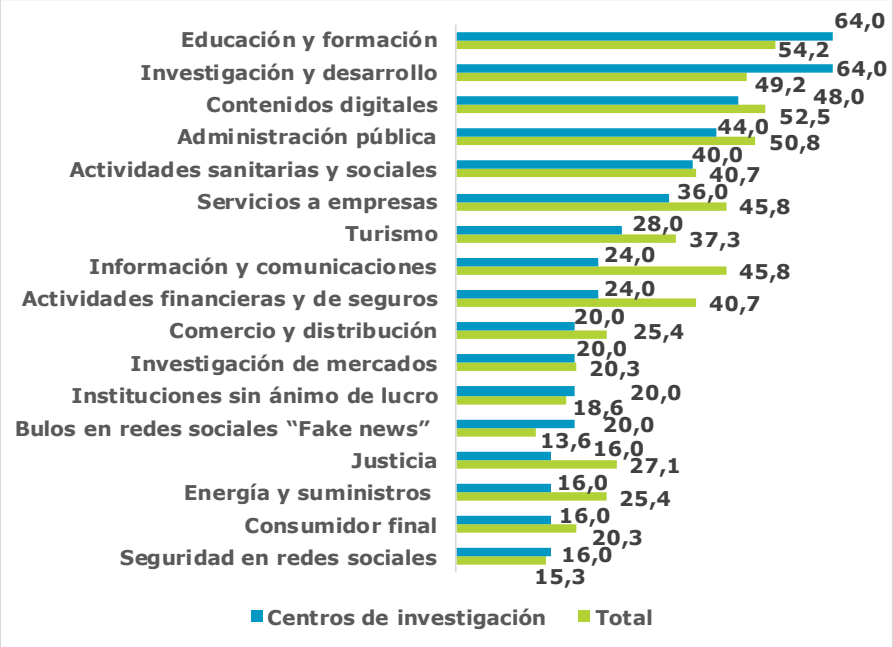
FIGURA 26. DESTINO FUNCIONAL DE LAS VENTAS DE TL DE LAS EMPRESAS %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 59, Empresas 34. P13

Por su parte, el sector de la Investigación y desarrollo y el sector de Educación y formación son a los que están más orientados los centros de investigación con un 64%.

FIGURA 27. DESTINO FUNCIONAL DE LAS VENTAS DE TECNOLOGÍA DEL LENGUAJE DE LOS CENTROS DE INVESTIGACIÓN %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 59, Centros de investigación 25. P13

La mayoría de las empresas afirmaron en las entrevistas en profundidad realizadas que el destino funcional de las ventas en el sector de las tecnologías del lenguaje es transversal, afecta a todo tipo de sectores porque no son actividades finalistas que vayan dirigidas a un uso particular o un cliente concreto. Sin embargo, aunque se trate de una tecnología transversal las empresas afirmaron que han de adaptarse al cliente al que vaya dirigida finalmente la solución, ya que cada sector de actividad tiene unos tecnicismos y unas particularidades en el lenguaje que han de ser consideradas:

"Simplemente hacemos un desarrollo de productos, de servicios, muy amplio adaptado a cada uno de los clientes, aunque haya muchos elementos comunes, pero en cada caso hay que particularizar porque cada cliente tiene necesidades particulares y adaptaciones, no es lo mismo el lenguaje que utilizamos para la banca, para los seguros, para la sanidad, etc.".

No obstante, coincidieron en afirmar que sus ventas van dirigidas a sectores con grandes volúmenes de consultas de usuarios finales, como por ejemplo la banca, las compañías de seguros, el sector sanitario y farmacéutico o los servicios de atención de emergencias 112: *"porque quieren automatizar*

y optimizar sus canales de atención al cliente". Las empresas consideran que las soluciones de tecnologías del lenguaje optimizan la interacción con sus clientes y les permiten un ahorro de costes asociados al mantenimiento de call center.

En las entrevistas en profundidad realizadas los centros de investigación destacaron:

El sector bancario por la demanda existente de mejorar la interacción de sus canales de atención al cliente y por la existencia de todo tipo de documentos que les interesa analizar para mejorar su competitividad, como hipotecas o normativas internas. En esta línea, los centros de investigación también mencionaron su trabajo con el sector turístico en tanto que parte de su actividad está fuertemente relacionada con la atención al cliente.

El sector sanitario y la investigación biomédica**,** donde la aplicación de las tecnologías del lenguaje a la información y opiniones vinculada a dosieres clínicos se ha relevado útil para la realización de diagnósticos y el apoyo a la decisión clínica. También destaca la intervención sanitaria en personas de la tercera edad a través de tecnologías de reconocimiento de la voz.

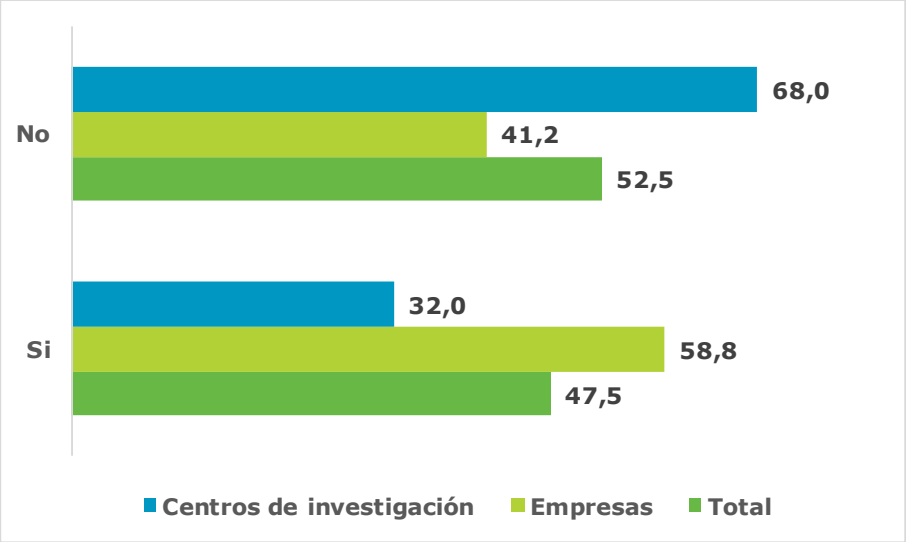
Los centros de investigación expresaron finalmente trabajar en proyectos de movilidad inteligente, en el ámbito de proyectos de ciudades y territorios inteligentes relacionados con la creación de puntos de información en la calle, centros de atención al ciudadano o centros conversacionales.

Internacionalización

En lo que respecta a la internacionalización del sector, el 52,5% de los agentes entrevistados ha indicado que no exporta productos o servicios relacionados con las tecnologías del lenguaje a otros países. No obstante, la situación es totalmente contraria si observamos el comportamiento de empresas y el de centros de investigación.

El 58,8% de las empresas exporta productos o servicios relacionados con las tecnologías del lenguaje a otros países, mientras tan solo el 32% de los centros de investigación consultados realiza exportaciones. Esto nos muestra que el tejido empresarial está más internacionalizado que el de centros de investigación, sin duda por su naturaleza comercial y por el ámbito de actuación limitado de los centros de investigación.

FIGURA 28. AGENTES QUE EXPORTAN PRODUCTOS/SERVICIOS RELACIONADOS CON LAS TECNOLOGÍAS DEL LENGUAJE A OTROS PAÍSES %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 59, Empresas 34, Centros de investigación 25. P13

En las entrevistas en profundidad realizadas, la mayoría de las empresas se situaron en dos posiciones opuestas: aquellas empresas que basan su mercado en el exterior y aquellas empresas que tienen todos sus clientes en el mercado español. En este sentido, se preguntó a las empresas entrevistadas sobre los factores clave en la internacionalización de las empresas del sector.

Por un lado, mencionaron la visibilidad en los mercados exteriores acompañada de presencia web:

"Visibilidad a través de ferias internacionales para que te conozcan y tener una marca que produzca confianza, básicamente visibilidad y buen servicio y no mucho más. No tenemos aranceles, no hay productos, no hay demasiado impedimento para exportar, es simplemente estar ahí, tener visibilidad".

Algunas empresas afirmaron que el sector de las tecnologías del lenguaje es internacional en sí mismo porque se basa en poder dar servicios en diferentes idiomas. Destacan, en algún caso, la publicidad como factor clave para conseguir un gran volumen de exportaciones.

Por otro lado, las empresas señalaron que se enfrentan al desafío de obtener grandes cantidades de recursos lingüísticos para entrenar sus sistemas en otro idioma:

"En el área del procesamiento del lenguaje natural el léxico es fundamental, ya que para ir a trabajar al país al que la empresa quiera abrir sus exportaciones, necesita que su plataforma pueda manejarse en ese idioma, para lo que necesita grandes cantidades de léxico".

Un grupo de empresas afirmaron que en el sector de las tecnologías del lenguaje hay tantos mercados como lenguas. También plantean que, para el mercado del idioma inglés, donde existe presencia de grandes multinacionales proveedores de contenidos, como pueden ser Google o Microsoft, se vuelve prácticamente imposible competir para una pequeña empresa española, por lo que opinan que la clave de la internacionalización de las empresas españolas es **abrirse al mercado de la lengua castellana**, esto es al mercado latinoamericano.

La mayoría de los centros de investigación expresaron en las entrevistas en profundidad mantener relaciones internacionales con otros centros de investigación europeos y americanos, con los que colaboran en el intercambio de profesionales, en el desarrollo de investigaciones y en publicaciones conjuntas. Además, participan en congresos y revistas internacionales, así como en proyectos europeos de investigación.

Los centros de investigación expresaron que el factor clave para su internacionalización es la colaboración en proyectos comunes al sector de tecnologías del lenguaje:

"La idea de tener una estrategia conjunta, porque a veces el sector está un poco atomizado. Hay que buscar ejes, talento, intentar sumar, es un factor importante para dar un salto a nivel internacional".

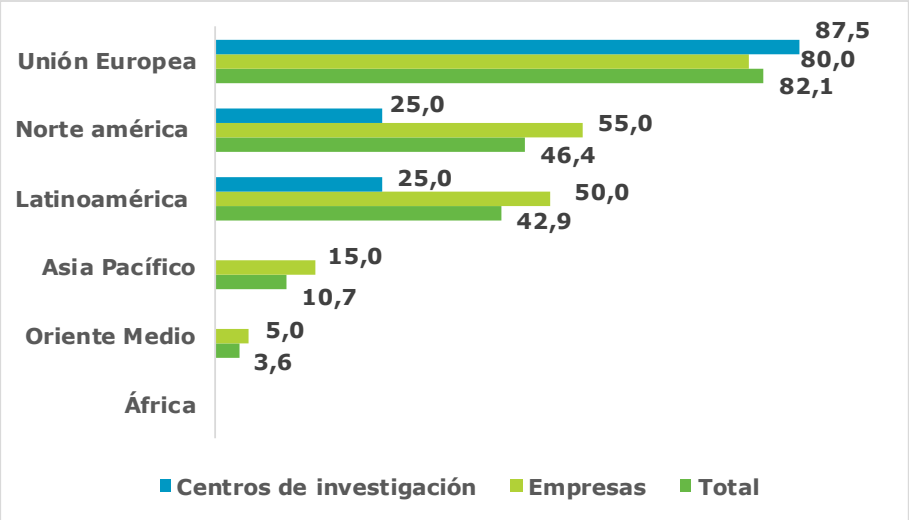
Ámbito geográfico de las ventas del sector

El ámbito geográfico al que van dirigidas en mayor proporción las ventas de los agentes consultados del sector es la Unión Europea, con un 82,1% de las ventas.

En segundo y tercer lugar se encuentra Norteamérica con un 46,4% y Latinoamérica con un 42,9%.

En el caso de las empresas consultadas del sector, el 55% dirigen sus ventas a Norteamérica y el 50% a Latinoamérica, mientras tan solo el 25% de los centros de investigación dirigen sus ventas a estas áreas geográficas. Esto mostraría que las empresas consultadas del sector están más abiertas a estos dos mercados que los centros de investigación, más orientados a participar en los proyectos financiados por los programas de I+D+i de la Unión Europea.

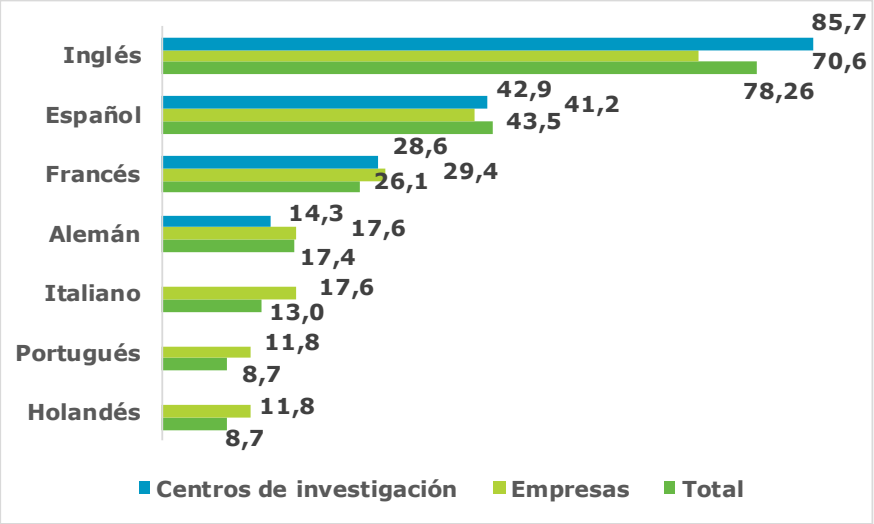
FIGURA 29. ÁMBITO GEOGRÁFICO DEL SECTOR %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 28, Empresas 20, Centros de investigación 8. P15b

Respecto a las lenguas en las que exportaron los productos o servicios de tecnologías del lenguaje los agentes que afirmaron que su ámbito geográfico se encontraba en la Unión Europea, la lengua predominante es el inglés, con un 78,2% de los agentes consultados. Cabe destacar, que los centros de investigación están un poco más orientados hacia la lengua inglesa en la Unión Europea que las empresas (un 85,7% de los centros frente a un 70,6% de las empresas).

FIGURA 30. LENGUAS EN LAS QUE LOS AGENTES EXPORTARON SUS PRODUCTOS/SERVICIOS A LA UNION EUROPEA %



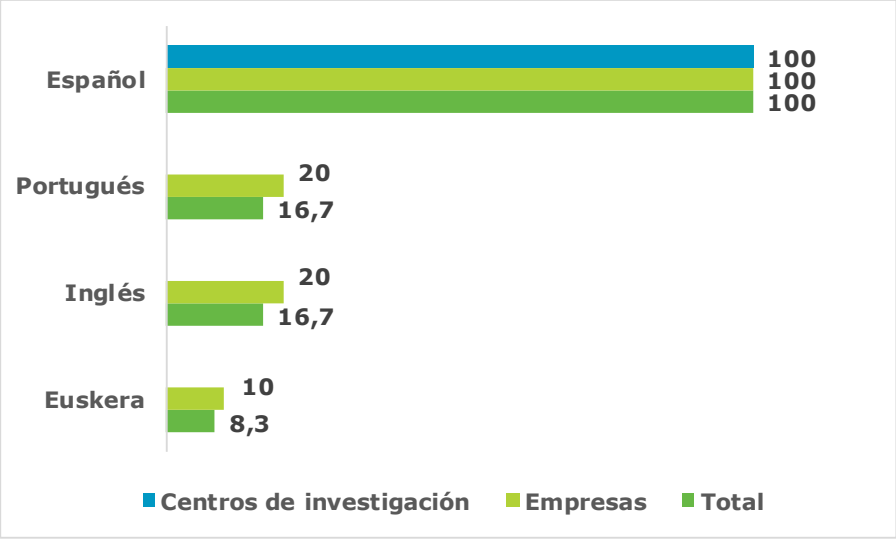
Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 23, Empresas 17, Centros de investigación 7. P15C

La segunda lengua hacia la que más orientan los agentes del sector consultados sus ventas en la Unión Europea es el español con un 43,5%, lo que podría mostrar el peso que tiene la lengua española en el mercado europeo.

Cabe señalar que alrededor del 4% de los agentes consultados afirmó dirigir sus servicios en lenguas minoritarias, entre las que se encuentra el árabe, ruso, noruego, danés, checo, sueco, croata, búlgaro, finés, griego, húngaro, letón, lituano, etc.

En esta categoría se encuentra también el catalán, el euskera y el gallego. Por otra parte, respecto a los agentes que dirigen sus ventas a Latinoamérica, el 100% lo hacen en español. También lo hacen en portugués y en inglés

FIGURA 31. LENGUAS EN LAS QUE LOS AGENTES EXPORTARON SUS PRODUCTOS/SERVICIOS A LATINOAMERICA %

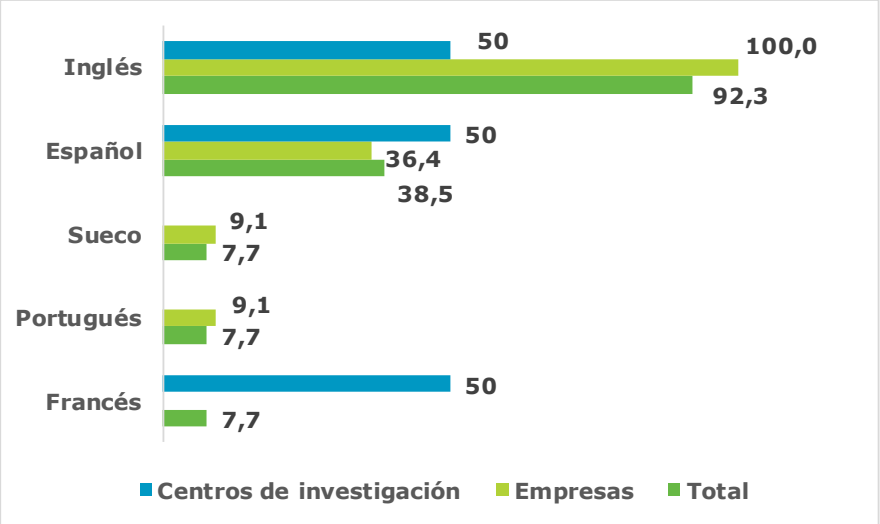


Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 12, Empresas 10, Centros de investigación 2. P15C

Los agentes del sector que dirigen sus ventas a Norteamérica lo hacen fundamentalmente en inglés (92,3%). En este sentido, las empresas consultadas del sector dirigen en la lengua inglesa el 100% de los productos y servicios que comercializan en Norteamérica.

Por otra parte, el 38,5% de los agentes del sector dirigen sus ventas hacia este continente en la lengua española. El resto de los idiomas tienen carácter testimonial.

FIGURA 32. LENGUAS EN LAS QUE LOS AGENTES EXPORTARON SUS PRODUCTOS/SERVICIOS A NORTEAMÉRICA %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 13, Empresas 11, Centros de investigación 2. P15C

El volumen de facturación que se corresponde con las actividades de tecnologías del lenguaje se situó entre 201 millones de euros y 211 millones de euros.

El 54,2% de los agentes del sector han aumentado su volumen de clientes durante el último año, lo que reforzaría la idea de un sector en auge.

Los agentes consultados dirigen sus ventas hacia más de 11 sectores de actividad distintos, lo que indica la transversalidad de la aplicación de las tecnologías del lenguaje a todo tipo de sectores de actividad. No obstante, destaca el sector de la educación y formación, contenidos digitales y sector público.

El sector de tecnologías del lenguaje es transversal a todo tipo de actividades productivas, las empresas señalaron como destino de sus ventas sectores con grandes volúmenes de consultas de usuarios finales, como por ejemplo la banca, compañías de seguros, el sector sanitario y farmacéutico o los servicios de atención de emergencias.

El tejido empresarial está más internacionalizado que los centros de investigación por su naturaleza comercial y por el limitado ámbito de actuación de los centros de investigación. Por otra parte, el ámbito geográfico de los agentes internacionalizados es, fundamentalmente, la Unión Europea, especialmente en el caso de los centros de investigación que, a través de los programas de financiación en I+D+i europeos, se orientan fundamentalmente hacia Europa.

3 El modelo de negocio

3.1 Cadena de valor

La cadena de valor del sector de tecnologías del lenguaje podría plantearse en dos fases: una primera fase preindustrial, que va desde el tratamiento de la materia prima hasta la aplicación de herramientas o componentes básicos a la información normalizada, y una segunda fase industrial, donde se desarrollan las aplicaciones que se comercializan en productos o servicios, listas para ser consumidas por ciudadanos, empresas, administraciones, y otras entidades.

A continuación, se describen las fases de la cadena de valor por las que pasa el desarrollo de una solución de tecnologías del lenguaje.

Fase preindustrial:

- **Recursos básicos**: los recursos básicos para las tecnologías del lenguaje están formados por la información y los textos sin tratar, es decir, la materia prima antes de ser puesta en valor. Es necesario tratar los recursos básicos para transformarlos en infraestructuras lingüísticas, esto es, clasificarlos, normalizarlos y computarlos.
- **Recursos o infraestructuras lingüísticas:** son el conjunto de información tratada, normalizada y/o anonimizada, transformada en un conjunto de corpus que se necesitan para entrenar los sistemas de tecnologías del lenguaje. Son los denominados recursos para las tecnologías del lenguaje en la clasificación de productos y soluciones TL:
    - Corpus de referencia, corpus sintácticos, corpus de discurso.
    - Corpus paralelo o memorias de traducción.
    - Lexicones.
    - Bases de datos terminológicas
    - Ontologías
- **Herramientas o componentes básicos:** en un tercer nivel encontramos las herramientas básicas para las que se necesita un conocimiento textual que proviene de las infraestructuras lingüísticas y la aplicación de la algoritmia. Estas herramientas básicas son las necesarias para el desarrollo de las aplicaciones finales. Algunos ejemplos de herramientas básicas:

- Herramientas de análisis y generación morfológica y desambiguación.
- Herramientas de análisis sintáctico, reconocimiento de entidades, semántica léxica y oracional.
- Herramientas de procesamiento avanzado de discurso.

En la fase preindustrial es donde se produce la investigación, dependiente en mayor medida de los grupos de investigación, como se verá más adelante en la descripción de los agentes involucrados en la cadena de valor. En estas fases de la cadena de valor es donde encaja el "open source" o código abierto en dos sentidos, por un lado, permite a los agentes acceder a recursos básicos para convertir en infraestructuras lingüísticas que permitan crear herramientas o componentes de tecnologías del lenguaje. Por otro lado, funciona como facilitador a la investigación en la medida que unas infraestructuras lingüísticas en código abierto permiten realizar pruebas de los algoritmos lingüísticos que los conviertan en herramientas o componentes básicos de tecnologías del lenguaje.

Fase industrial:

- **Desarrollo de aplicaciones:** en la fase industrial se encuentra el desarrollo de aplicaciones finales listas para comercializar. Las aplicaciones incorporan herramientas y componentes básicos de tecnologías del lenguaje y la algoritmia. En la fase industrial es donde se produce el desarrollo de aplicaciones y la innovación en la creación de nuevas aplicaciones. Algunos ejemplos de aplicaciones:
  - Sistemas de traducción automática y ayuda a la traducción.
  - Sistemas de dialogo.
  - Sistemas de extracción de información y text analytics.
  - Aplicaciones de ayuda a la redacción.
  - Aplicaciones de generación de textos.

La comercialización de las soluciones desarrolladas se puede realizar en dos sentidos:

- Por un lado, el desarrollo de aplicaciones se puede comercializar en forma de productos que venden las empresas desde sus instalaciones y servidores. Los productos de tecnologías del lenguaje se comercializan por medio de la venta de licencias, aunque el soporte de las licencias y el mantenimiento de los productos ha de realizarse desde las instalaciones de las empresas.

Esto implica que han de tener capacidad para dar respuesta a los clientes en el funcionamiento de sus productos, así como en su mantenimiento, lo que no suele estar al alcance de las pymes españolas. Las grandes empresas norteamericanas se encuentran fundamentalmente en la comercialización de productos, ya que tienen la capacidad para crear un producto acabado, mantenerlo y comercializarlo.

- El otro camino que pueden seguir las empresas del sector para comercializar sus soluciones es ofreciendo sus servicios a través de un modelo de prestación de servicios. En este caso, las pymes españolas ofrecer un servicio muy concreto de tecnologías del lenguaje que no precisa la inversión en tecnología y mantenimiento que requieren las licencias de productos, en este sentido, les permite ahorrar costes asociados a mantenimiento de infraestructuras. La venta de servicios a través de licencias es más personalizada que la de productos y ha de adaptarse al cliente.

En la siguiente figura, se muestra gráficamente la cadena de valor desde la fase preindustrial hasta la fase industrial. Cabe señalar que, en cierta medida, hay que tener en cuenta que la cadena de valor del sector del lenguaje se retroalimenta de manera que, con las infraestructuras lingüísticas se pueden hacer las herramientas básicas y con las herramientas se desarrollan las aplicaciones, pero también con las herramientas básicas puedes crear nuevas infraestructuras lingüísticas, y con el desarrollo de aplicaciones puedes crear nuevas herramientas básicas.

En este sentido, infraestructuras lingüísticas, herramientas básicas y aplicaciones se encontrarían en un círculo de retroalimentación.

FIGURA 33. CADENA DE VALOR DEL SECTOR TECNOLOGÍAS DEL LENGUAJE

Respecto a la intervención de los principales agentes identificados en la cadena de valor de tecnologías del lenguaje:

- **Los grupos de investigación** se encuentran en la fase preindustrial como facilitadores de infraestructuras lingüísticas y herramientas del lenguaje. Los grupos de investigación no tienen como fin la comercialización de las soluciones sino la investigación, por lo que son la base necesaria para la generación de herramientas básicas de tecnologías del lenguaje y la investigación en la fase preindustrial. Los grupos de investigación no están orientados al mercado y en general no tienen los recursos económicos ni de personal que requiere el desarrollo de aplicaciones. No obstante, a través de la creación de empresas "*spin-off*" los grupos de investigación pueden dar el salto comercial con alguna solución concreta en la que se hayan especializado y con la que decidan salir al mercado.

- **Las empresas** están orientadas a la fase industrial por su propia naturaleza. Son empresas muy especializadas y con pocos trabajadores, en su mayoría microempresas, que están presentes en la fase de desarrollo de aplicaciones y consumen de otras empresas o centros de investigación las infraestructuras y herramientas necesarias para el desarrollo de sus soluciones finales.

  Las empresas con cierto tamaño o recursos generan sus propias infraestructuras lingüísticas a partir de recursos básicos, sobre las que aplican herramientas básicas de tecnologías del lenguaje que más tarde desarrollan en aplicaciones. Estas infraestructuras lingüísticas suelen pertenecer a sus clientes por lo que solo pueden utilizarlas en determinados contextos y con el permiso de sus clientes. En este sentido, algunas empresas se encuentran en toda la cadena de valor.

- Las **administraciones públicas** también pueden promover, apoyar o intervenir en todas las fases de la cadena de valor.

  En primer lugar, en la puesta a disposición y la apertura de los recursos básicos lingüísticos con los que cuentan en todos los ámbitos, y en particular el sanitario o el de justicia, para su transformación en infraestructuras lingüísticas.

  En segundo lugar, la administración pública puede intervenir en la publicación de convocatorias para aplicar tanto herramientas básicas de tecnologías del lenguaje como aplicaciones finales a su funcionamiento, procesos de gestión y servicios de atención al ciudadano.

La solución comercializada está dirigida a dos tipos de clientes, por un lado, a un cliente directo o final, es aquel que ha comprado el producto o servicio de tecnologías del lenguaje, pudiendo ser una empresa, una administración, una institución, organización o asociación, y por otro lado, a un cliente indirecto, que es el cliente que se beneficia de manera implícita de la tecnología del lenguaje aplicada, este tipo de clientes suele ser la sociedad en su conjunto o los individuos particulares.

Por ejemplo, cuando una persona en Dinamarca compra un producto de una empresa textil española desde la página web, puede traducirla a su idioma para poder navegar por el catálogo de productos y entender la información, por lo que en este caso el cliente se beneficia de la aplicación de la traducción automática a la página web de la empresa textil española.

3.2 Modelo de ingresos

El modelo de ingresos de los agentes del sector se ha analizado desde la óptica de la venta de servicios, la venta de licencias, el acceso a subvenciones y la modalidad gratuita de prestación de servicios.

Por un lado, la venta de servicios de tecnologías del lenguaje se realiza mediante:

- Venta de servicios profesionales para el cliente final por trabajo realizado.
- Venta de servicios profesionales mediante redes de partners.
- Venta de servicios profesionales mediante acuerdos comerciales.

Por otro lado, la venta de licencias de productos y/o servicios de tecnologías del lenguaje se realiza mediante:

- Pago por licencias asociadas al pago por uso/acceso/trabajo realizado.
- Licencias de cesión de productos y servicios de tecnologías del lenguaje.
- Licencias recíprocas o copyleft: GPL y LGPL.
- Licencias permisivas: MIT, BSD, Apache, MPL y EPL.

Para terminar, las empresas y los centros de investigación del sector pueden comercializar sus productos y/o servicios a través del acceso a subvenciones de I+D+i y mediante el acceso a servicios gratuitos sin restricciones.

El modelo de ingresos que más utilizan los agentes del sector consultados para comercializar sus soluciones es la venta de servicios profesionales para el cliente final por trabajo realizado con un 72,4%.

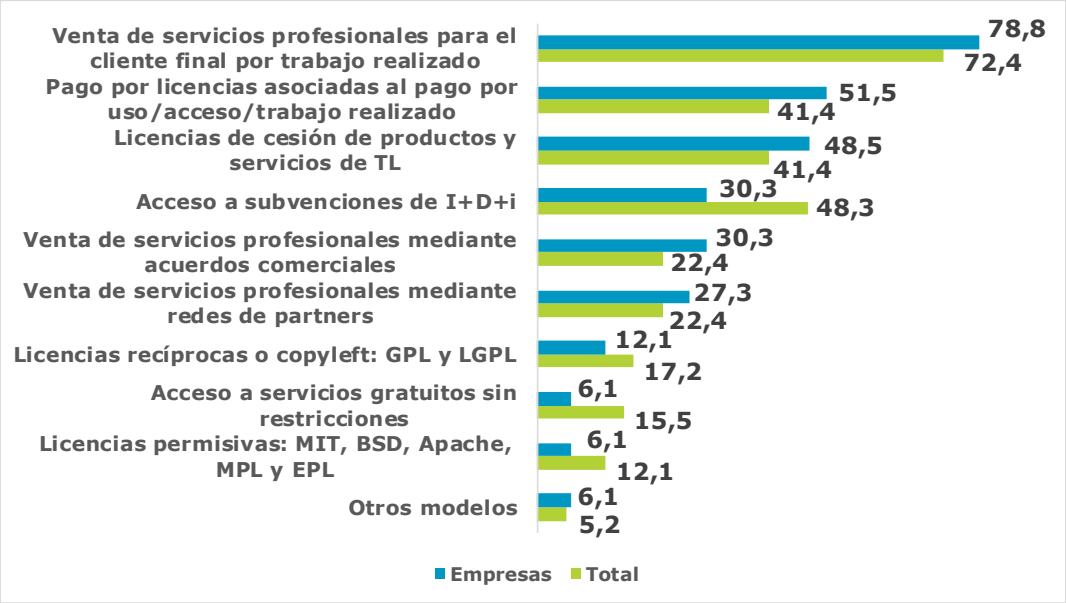
En segundo lugar, un 48,3% de los agentes consultados manifestó que su modelo de ingresos se basa en el acceso a subvenciones de I+D+i. En tercer lugar, un 41,4% de los agentes consultados comercializan sus productos y servicios a través de licencias asociadas a uso/acceso/trabajo realizado y licencias de cesión de productos y servicios de tecnologías del lenguaje.

Respecto a las empresas, el modelo de ingresos más utilizado es la venta de servicios profesionales para el cliente final por trabajo realizado con un 78,8%.

En segundo lugar, el 51,5% de las empresas comercializa sus productos o servicios a través del pago por licencias asociadas al pago por uso, acceso y/o trabajo realizado.

En tercer lugar, un 48,5% de las empresas comercializa a través de licencias de cesión de productos y servicios de tecnologías del lenguaje.

FIGURA 34. MODELO DE INGRESOS EMPRESAS DEL SECTOR %

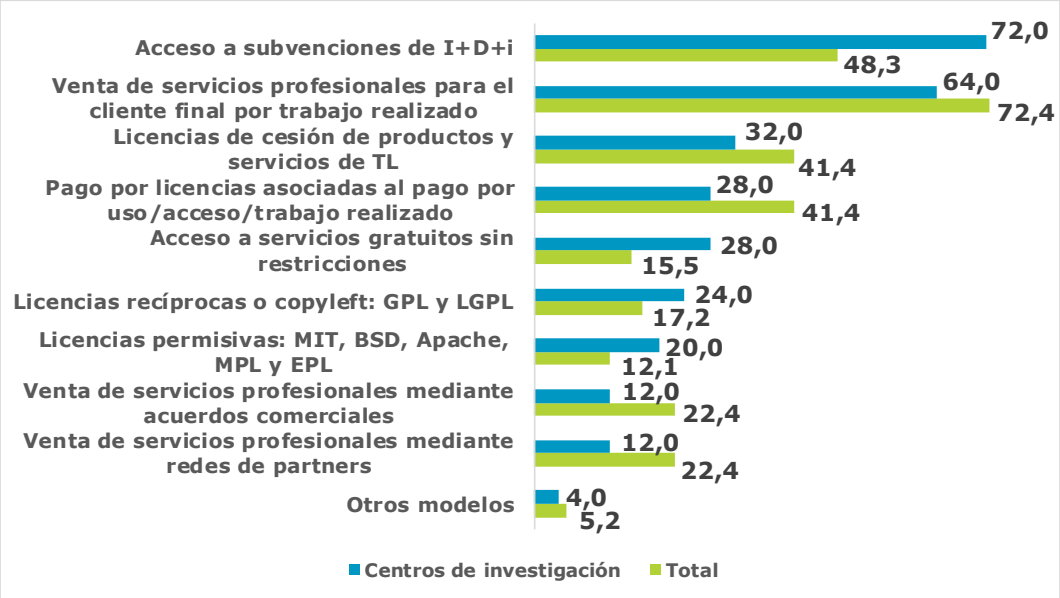


Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 58, Empresas 33. P21

Por otro lado, el modelo de ingresos que más utilizan los centros de investigación consultados es el acceso a subvenciones de I+D+i con un 72%. En segundo lugar, el 64% de los centros de investigación comercializa mediante la venta de servicios profesionales para el cliente final por trabajo realizado.

En tercer lugar, el 32% de los centros de investigación consultados utilizan licencias de cesión de productos y servicios de tecnologías del lenguaje.

FIGURA 35. MODELO DE INGRESOS CENTROS DE INVESTIGACIÓN DEL SECTOR %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 58, Centros de investigación 25. P21

Tipos de clientes y modalidad de transferencia

Los agentes del sector de tecnologías del lenguaje utilizan cuatro modalidades de transferencias para comercializar sus productos y servicios con sus clientes.

Por un lado, los agentes del sector pueden transferir las aplicaciones que crean en la modalidad de software de código abierto, este software permite que el código fuente y otros derechos que normalmente son exclusivos para quienes poseen los derechos de autor, sean publicados bajo una licencia de código abierto y formen parte del dominio público.

Esto permite a los clientes utilizar, cambiar y redistribuir el software, ya sea en su forma modificada o en su forma original. Por otra parte, los agentes del sector pueden transferir productos acabados a sus clientes o transferir componentes de sus productos o servicios, partes atomizadas y específicas que pueden comercializar. Y en último lugar, los agentes del sector pueden transferir a sus clientes asesoramiento y/o formación en el sector de tecnologías del lenguaje.

Los agentes del sector dirigen la venta de sus productos y servicios de tecnologías del lenguaje a la sociedad en general o a los ciudadanos como cliente indirecto de los productos y soluciones que desarrollan. Por otra parte, pueden comercializar sus productos a las empresas, a centros de investigación, a la universidad, a instituciones sin ánimo de lucro, a asociaciones y a la administración.

La modalidad de transferencia de software de código abierto es más utilizada por los agentes del sector para dirigirla a la sociedad o a los ciudadanos como clientes indirectos (14,4%), a los centros de investigación y la universidad pública, con un 12,1% respectivamente.

El software de código abierto es un instrumento que permite a los centros de investigación y a las universidades investigar, desarrollar y compartir soluciones de tecnologías del lenguaje a las que no tendrían acceso de otra forma por falta de recursos para adquirir licencias e integrarlas en sus plataformas.

El producto o servicio acabado de los agentes del sector consultados va dirigido en mayor medida a empresas nacionales (16,2%) y empresas internacionales (13,6%).

Los componentes de productos o servicios van dirigidos de nuevo en mayor medida a empresas nacionales (14,2%), centros de investigación (11,2%) y empresas internacionales (10,5%). Lo que podría mostrar la retroalimentación de la cadena de valor del sector, donde empresas y centros de investigación destinan la producción o desarrollo de servicios a las propias empresas y centros de investigación del sector y, además, a empresas internacionales, como se verá más adelante en el apartado 4.1 Cadena de valor del sector.

El asesoramiento y formación de los agentes del sector consultados va dirigido, fundamentalmente, a empresas nacionales (15,1%) y a la Universidad pública (11,8%).

Cabe señalar que el 12% de los agentes del sector consultados no utiliza ninguna de las modalidades de transferencia descritas para relacionarse con la administración europea y la administración internacional, lo que podría señalar que utilizan otro tipo de transferencia con estas administraciones.

TABLA 2: MODALIDAD DE TRANSFERENCIA POR TIPO DE CLIENTE AL QUE VA DIRIGIDA %

Tipos de clientes	Modalidad de transferencia				
	Software de código abierto	Producto acabado	Componente de producto	Asesoramiento y/o formación	Ninguna
Sociedad	14,4	7,66	5,97	9,14	6,85
Empresas nacionales	8,08	16,17	14,18	15,05	1,64
Empresas europeas	5,05	9,79	9,7	9,14	7,12
Empresas internacionales	5,05	13,62	10,45	7,53	5,48
Centros de investigación	12,12	4,26	11,19	9,14	6,03
Universidad pública	12,12	7,23	7,46	11,83	5,48
Universidad privada	5,05	5,11	5,22	4,84	9,86
Instituciones sin ánimo de lucro	6,06	3,83	6,72	7,53	9,04
Asociaciones	7,07	5,53	4,48	5,91	9,32
Adminstración autonómica	9,09	9,36	9,7	8,06	7,12
Administración nacional	8,08	9,79	6,72	6,99	7,67
Administración europea	4,04	3,83	3,73	3,23	12,33
Administración internacional	4,04	3,83	4,48	1,61	12,05
Total	100	100	100	100	100

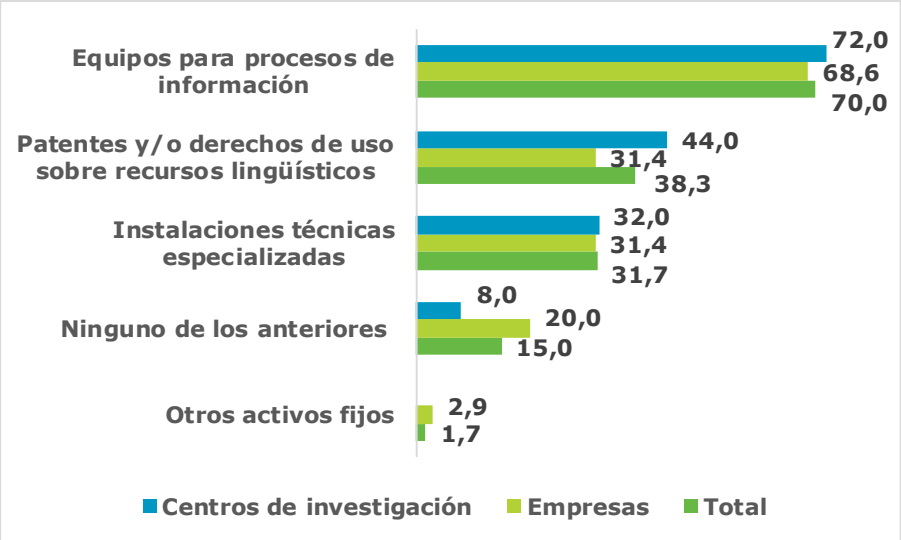
Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 59. P11

3.3 Modelo de producción

Activos fijos del sector

Los activos fijos más utilizados por los agentes del sector destacan en primer lugar los equipos para procesos de información con un 70%, seguido de patentes y/o derechos de uso sobre recursos lingüísticos con un 38,3%, lo que guarda cierta lógica ya que el 61,7% de los agentes manifestó dedicar su actividad a desarrollar recursos para las tecnologías del lenguaje.

FIGURA 36. ACTIVOS FIJOS UTILIZADOS POR LOS AGENTES DEL SECTOR %



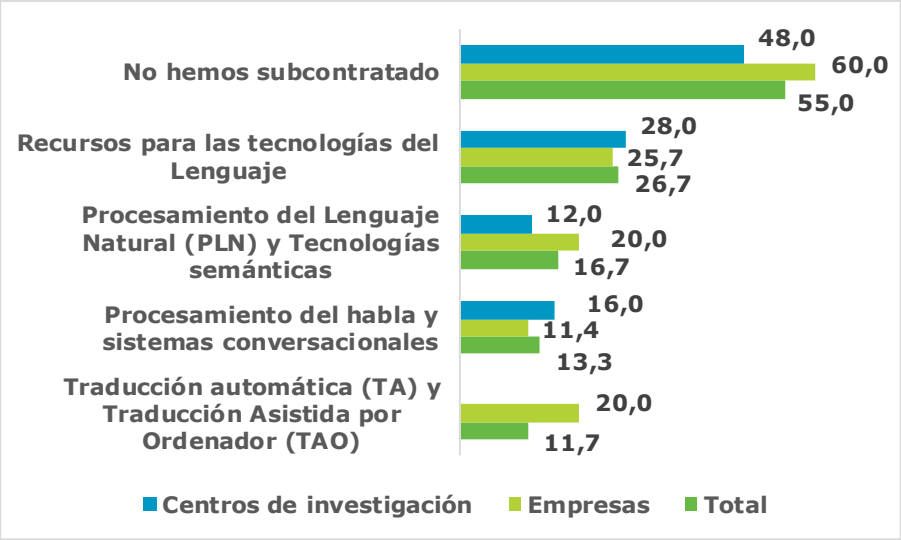
Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 58, Empresas 33, Centros de investigación 25. P9

Un 20% de las empresas consultadas no mantiene ninguno de estos activos fijos frente a un 8% de los centros de investigación, cabe señalar que los ordenadores forman parte de los equipos para procesos de información, por lo que es posible que la interpretación de esta pregunta no haya sido la adecuada por parte de las empresas. No obstante, podría señalar que el modelo de producción de las empresas consultadas está más orientado a ofrecer servicios para los que no precisan de activos fijos, si contrastamos este porcentaje de empresas que han señalado no utilizar ningún activo fijo con la mayoría de las empresas que comercializa a través de la venta de servicios profesionales para el cliente final por trabajo realizado.

Contratación a terceros de servicios del sector

Poco más de la mitad de los agentes consultados (55%) no ha subcontratado servicios relacionados con las tecnologías del lenguaje a terceros, aumentando hasta un 60% si atendemos a las empresas. Al tratarse de empresas que, en su gran mayoría comercializan con soluciones de tecnologías del lenguaje de todo tipo, es decir, bastante heterogéneas en el tipo de soluciones que desarrollan, podrían no necesitar subcontratar ningún producto o servicio específico. Los agentes consultados subcontratan en mayor medida recursos para las tecnologías del lenguaje (26,7%) que otro tipo de servicios.

FIGURA 37. PRODUCTOS Y SERVICIOS QUE SUBCONTRATAN LOS AGENTES DEL SECTOR %



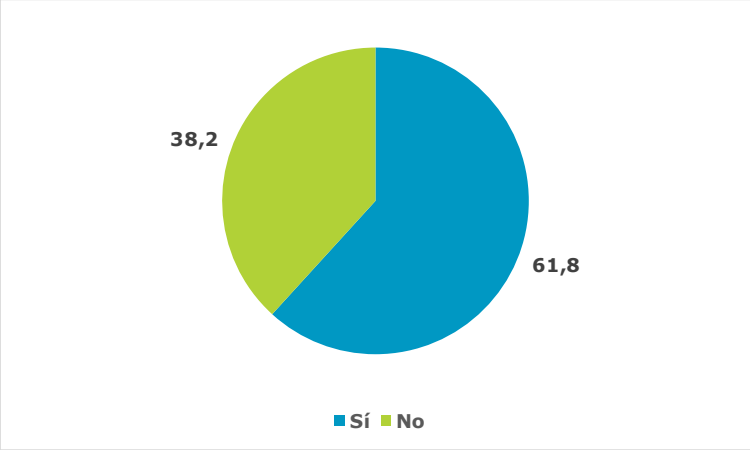
Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 58, Empresas 33, Centros de investigación 25. P10

3.4 Investigación e innovación en el sector

Inversión en I+D+i

La mayoría de las empresas consultadas (61,8%) manifestó que tienen un departamento de I+D+i para apoyar el desarrollo de soluciones de su negocio, lo que podría indicar el grado de innovación que implica el desarrollo de soluciones de tecnologías del lenguaje.

FIGURA 38. EMPRESAS CON DEPARTAMENTO DE I+D+I%



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 59, Empresas 34. P10

El volumen de inversión en I+D+i de las 16 empresas que contestaron a esta pregunta se sitúa en un total de 46 millones de euros en 2017. Esto supone que el índice de inversión en I+D+i de las empresas consultadas, representa el 6,3% de su facturación total. Si atendemos al volumen de inversión que se corresponde con actividades del sector de tecnologías del lenguaje, representa el 10,1% del volumen de su facturación que se corresponde con la actividad.

En las entrevistas en profundidad realizadas la mayoría de las empresas consultadas que comercializa con productos expresó que su empresa cuenta con departamento de I+D+i o, en el caso de empresas más pequeñas, empleados encargados del área de innovación.

Por otro lado, un grupo de empresas expresó que no tienen departamento de innovación porque no se dedican a desarrollar soluciones, sino que ofrecen servicios a sus clientes con software desarrollado por otras empresas. Para ellas no justifica formar un departamento de innovación, con sus consecuentes gastos fijos asociados, al ser solo proveedores de servicios lingüísticos.

Por otra parte, respecto a la valoración de la inversión en I+D+i en el sector, se detectaron en las entrevistas realizadas dos posiciones contrapuestas:

Por un lado, un grupo de empresas calificó la inversión en I+D+i en el sector como escasa, tanto por parte de las propias empresas como por parte de la Administración. En este sentido, señalaron el papel que tienen las instituciones públicas para ofrecer incentivos a la innovación a través de la compra pública innovadora o la publicación de convocatorias de ayudas que potencien el desarrollo de herramientas innovadoras que lleguen a ser competitivas en el mercado exterior. Algunas empresas señalaron que la baja inversión en I+D+i no es un hecho característico del sector de tecnologías del lenguaje, sino que es un problema estructural y transversal a todos los sectores de actividad a nivel nacional.

Por otro lado, un grupo de empresas afirmaron que si existe inversión en I+D+i en el sector, ya que forma parte de su naturaleza y es un factor clave para su competitividad. No obstante, señalaron que el problema se encuentra en el tamaño de las empresas que conforman el sector de las tecnologías del lenguaje en España, ya que no pueden entrar a competir con las empresas americanas por las grandes operaciones con inversores internacionales.

"Las empresas españolas del sector son pequeñas y nos cuesta mucho desarrollar nuestro propio pulmón, pero es que nos come la actividad del día a día, es un problema de tamaño de las empresas, somos muy pequeños".

"En esta batalla las empresas tienen que estar constantemente repuntándose y la innovación es fundamental, no es opcional, es que si no tienes en este sector estás muerto".

Además, algunas empresas señalaron que podría estar produciéndose una inversión atomizada, consecuencia de la falta de colaboración entre empresas y centros de investigación, por lo que podría haber empresas y centros invirtiendo en desarrollar las mismas soluciones.

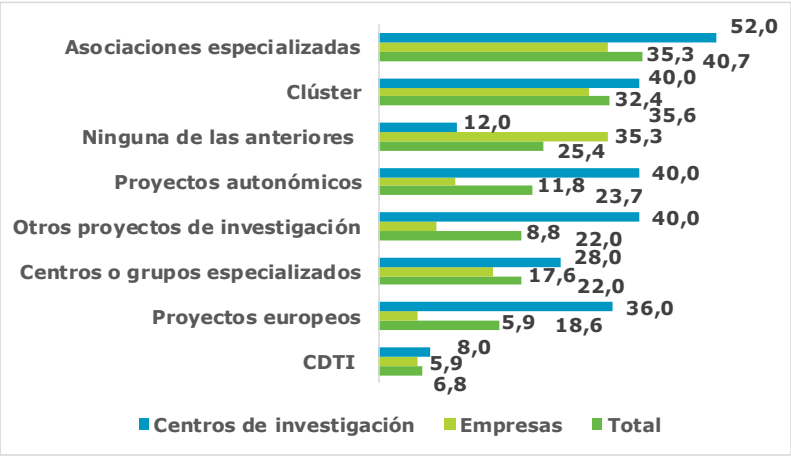
Desde la óptica de los centros de investigación, actúan ocasionalmente como el propio departamento de I+D+i de las empresas, fundamentalmente cediéndoles las herramientas básicas de tecnologías del lenguaje que crean y comprobando que las soluciones que las empresas desarrollan funcionan adecuadamente.

Por otro lado, coinciden en señalar que la inversión en el sector desde el punto de vista de convocatorias públicas ha ido disminuyendo de forma relevante durante los últimos años.

Redes de conocimiento del sector

Respecto a las redes de conocimiento en las que participan los agentes consultados del sector, destaca la participación en asociaciones especializadas (40,7%). Entre las asociaciones especializadas los agentes mencionaron Langune (16,7%) en primer lugar, y, en segundo lugar, la SEPLN y la RTTH (12,5% respectivamente). Otras asociaciones nombradas por empresas y centros del sector son AESLA, AETER, REALITER, RITERM, Big Data Value Asociation, ELRA, European Speech Communication, RSTDA, LT-Innovate, la Plataforma del español, AERFAI, Tekom Europa y CEEI Valencia.

FIGURA 39. REDES DE CONOCIMIENTO EN LAS QUE ESTÁN INTEGRADOS LOS AGENTES DEL SECTOR %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 59, Empresas 34, Centros de investigación 25. P19

En segundo lugar, el 35,6% de los agentes consultados participa en **clústers**. Como, por ejemplo, Eiken Cluster y Langune con un 28,6%, Gaia y Clusterlingua y Plataforma del Español (21,4%) y Madrid Network (14,3%).

Por otra parte, el 23,7% de los agentes que participan en proyectos de investigación autonómicos, como por ejemplo Berbaloa, Modela y Guales.

Entre los agentes que participan en otros proyectos (22%) han mencionado proyectos como AMIC-MINNECO, Genio Vox, CervanTIC, y Mecin.

Por último, los agentes que participan en proyectos europeos (18,6%) han mencionado proyectos como *Beaware* (European Commision), proyecto que consiste en incrementar la participación de los países de Europa oriental en actividades de investigación paneuropeas en materia de aeronáutica y transporte aéreo auspiciadas por Horizonte de 2020. Otro proyecto europeo mencionado es *Tensor* (European Commision , s.f.), relacionado con los retos a los que se enfrentan los países europeos en la identificación, recopilación e interpretación de los contenidos generados por el terrorismo en línea.

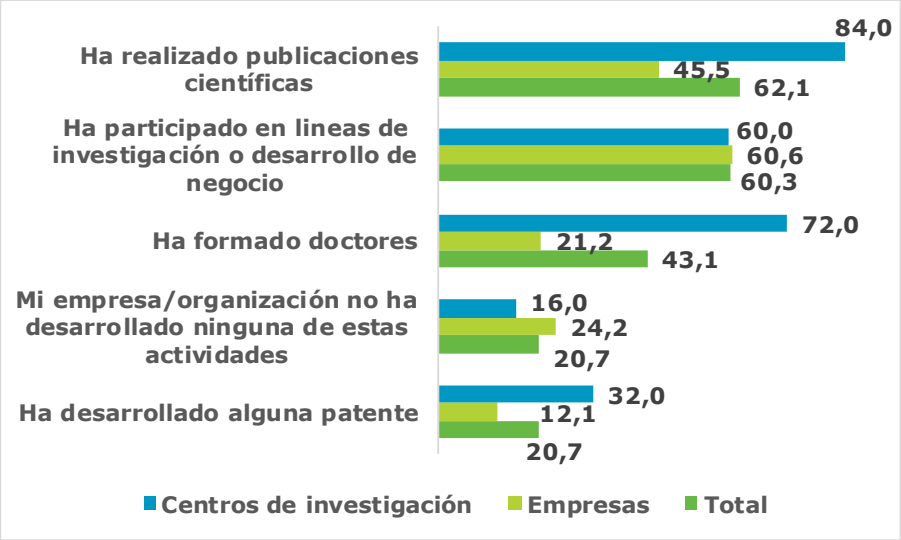
Los agentes también mencionaron el proyecto *Kristina* (Kristina, s.f.), proyecto europeo integrado en el Programa HORIZON 2020 de Investigación y Desarrollo de la Unión Europea. El objetivo que persigue es el de derribar las barreras lingüísticas existentes en la sociedad mediante el uso de tecnologías innovadoras para que los colectivos de inmigrantes y todos los grupos con problemas a la hora de comunicarse puedan acceder a una fuente de consulta médica interactiva a través de la red.

Cabe señalar que los centros de investigación participan en mayor medida que las empresas en las redes de conocimiento descritas, lo que guarda lógica con la naturaleza de los centros, basados en la investigación. Hasta un 35,3% de las empresas consultadas manifestó que no participa en ninguna red de conocimiento.

Líneas de investigación o desarrollo de negocio innovadoras

Tan solo un 20,7% de los agentes consultados manifestó que no ha desarrollado ninguna línea, lo que muestra el carácter innovador de este tipo de tecnologías.

FIGURA 40. LÍNEAS DE INVESTIGACIÓN O DESARROLLO DE NEGOCIO INNOVADORAS EN LAS QUE HAN PARTICIPADO LOS AGENTES %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 58, Empresas 33, Centros de investigación 25. P20

El 62,1% de los agentes consultados ha realizado publicaciones científicas y el 60,3% ha participado en líneas de investigación o desarrollo de negocio con un 60,3%.

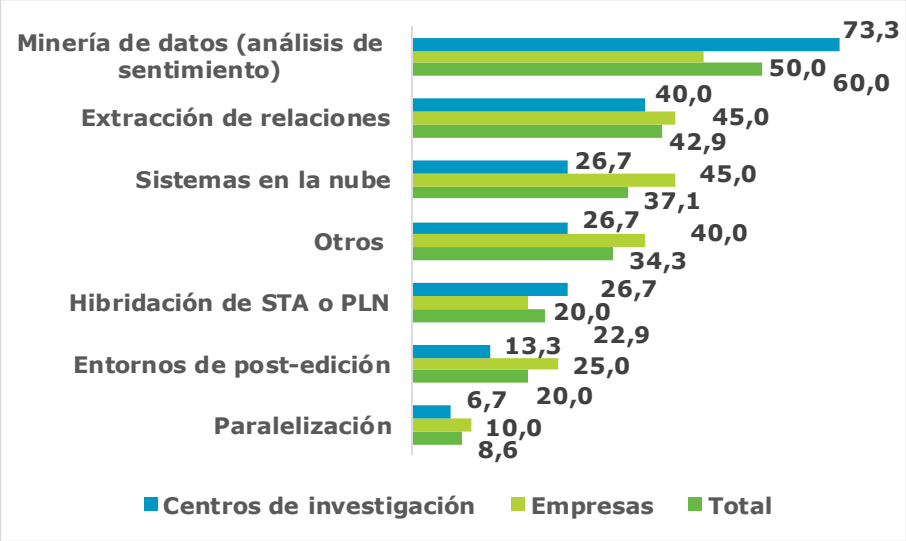
Cabe señalar que la mayoría de los centros de investigación (84%) ha realizado publicaciones científicas, mientras las empresas están más enfocadas a la participación en líneas de investigación (60,6%).

Destaca el papel que tienen los centros de investigación en la formación de doctores con un 72%. Por otra parte, tan solo el 20,7% de los agentes del sector ha desarrollado alguna patente.

Para terminar, se pidió a los agentes consultados que expresaron que participan en líneas de investigación que indicaran qué tipo de líneas abordan en mayor medida en sus negocios y organizaciones. En este sentido, la línea de investigación más desarrollada por los agentes es la minería de datos (análisis de sentimiento) con un 60%. Cabe señalar que el 73,3% de los centros de investigación consultados desarrollan líneas de investigación basadas en la minería de datos, lo que podría mostrar una tendencia de investigación en el sector de las tecnologías del lenguaje, ya que los centros son los que reproducen las líneas de investigación prioritarias que se establecen en Europa.

Por otra parte, la segunda línea de investigación o desarrollo de negocio en la que más participan los agentes consultados es la extracción de relaciones con un 42,9%.

FIGURA 41. LÍNEAS DE INVESTIGACIÓN O DESARROLLO DE NEGOCIO INNOVADORAS QUE DESARROLLAN LOS AGENTES %



Fuente: Encuesta del ONTSI Caracterización del sector de tecnologías del lenguaje 2017. Base Total 35, Empresas 19, Centros de investigación 16. P20A

La mayoría de las empresas consultadas en las entrevistas en profundidad afirmaron que en la actualidad mantienen líneas de investigación abiertas, algunas de las más nombradas fueron:

- Traducción automática (memorias de traducción).
- Desarrollo de analizadores morfológicos para edición y análisis de datos.
- Análisis de sentimiento.
- Componentes del procesamiento del lenguaje natural.
- Sistemas de lenguaje cognitivo.
- Algunas empresas, mantienen líneas de investigación de simulación empresarial, inteligencia artificial o redes neuronales.

Por su parte, los centros de investigación tienen líneas de investigación relacionadas con la combinación de Big Data y Big Analytics aplicado a los textos monolingües, es decir, recopilarlos con técnicas de Big Data y clasificarlos con técnicas de Big Analytics.

Por otra parte, tienen líneas de investigación relacionadas con el realce de voz, traducción y transmisión de voz, codificación y comprensión de la voz, recepción robusta, identificación de errores de transmisión, etc. También se identificaron líneas relacionadas con la tecnología de text to speech, procesamiento del lenguaje y procesamiento de la señal, simulación numérica, expresividad, interacción de robots y biometría.

Respecto al desarrollo de patentes, la práctica totalidad de las empresas consultadas no han desarrollado ninguna patente. Las empresas expresaron que no pueden permitirse el desarrollo de patentes por ser muy costoso para las empresas pequeñas que, además, no cuentan con tiempo para invertir en el desarrollo de nuevas patentes.

Una empresa apuntó que no se dedica al desarrollo de patentes dado que el software pertenece casi en su totalidad a las empresas norteamericanas, por lo que no le resulta rentable. En esta línea, algunas empresas argumentaron que trabajan con software libre:

"Tenemos una característica un poco especial y es que trabajamos con software libre, con lo cual huimos de un modelo basado en patentes, lo que queremos es compartir conocimiento y avanzar a partir de lo ya desarrollado y elaborado, por lo que perseguimos liberar".

Algunas empresas expresaron que se trata de un factor cultural:

"En otros sectores veo que hay esa cultura de querer publicar a través de patentes, pero en el ámbito del desarrollo de software en el que nos movemos nosotros la cultura es casi toda la contraria, no hay cultura de patentar".

Otros centros de investigación argumentaron que en Europa no se permite patentar este tipo de software, por lo que la forma de patentar es a través del registro de la propiedad intelectual.

"En la Universidad como indicador de calidad e investigación se usan las patentes, pero como lo que se produce es software y el software en Europa no se puede patentar, no patentamos. En Estados Unidos sí se puede patentar, entonces como utilizamos dinero público lo que hacemos es software libre y el problema es que eso no se valora en los indicadores de calidad de la I+D+i".

En este sentido, habría una competencia desigual entre Europa y EEUU en el desarrollo de patentes.

Colaboración en el sector

En general, las empresas y centros de investigación afirmaron en su mayoría colaborar entre ellas, ya sea por medio de acuerdos entre empresas, universidades, asociaciones o centros tecnológicos. Los acuerdos entre empresas se dan orientados a la comercialización concreta de productos o en el desarrollo de nuevas tecnologías.

Los centros de investigación expresaron que colaboran entre ellos en proyectos comunes o en el desarrollo de líneas de investigación, y colaboran con las empresas actuando como departamento de I+D+i, como consultores o desarrolladores, cediéndoles personal formado o realizando asesorías tecnológicas a las empresas.

En uno de los grupos de discusión que se realizó, se habló de un posible modelo de colaboración del sector:

"Podríamos llamarlo "coompetición", es decir, cooperamos y competimos en algunos casos. Es un buen modelo, creo que hay posibilidades de colaboración porque la competencia no es tan frontal, casi siempre hay posibilidades de colaboración, aunque no existen redes explícitas donde nos encontremos para hablar del negocio".

"Sí, nosotros competimos con otras empresas, pero también cooperamos, eso para una empresa de traducción dedicada al mundo de Internet es muy habitual, por ejemplo, estos señores cuando instalan una gestión de contenidos muy sofisticada a un cliente y tiene que estar en diecisiete idiomas y actualizarse cada día, pues tienen que cooperar con empresas como nosotros".

"Yo creo que hay dos modelos que sí pueden funcionar, uno de ellos es entre empresas que compiten directamente, muy difícil hacer innovación en lo que es la solución final pero sí en productos preindustriales, por ejemplo, las nuevas tecnologías en lo que son las pruebas de traducción automática neuronal, pueden juntarse varias empresas para capacitar a personal en su uso, o para hacer modelos de entrenamiento o de integración, sin llegar a un producto final. Y luego hay otro modelo que funciona muy bien, que es entre empresas de distintas actividades, se pueden hacer cosas de valor añadido y no competimos entre nosotros".

En resumen, en el sector de tecnologías del lenguaje se produce una suerte de "coompetición" entre empresas en dos sentidos:

- Entre empresas que se dedican a la misma actividad en el desarrollo de productos preindustriales.

- Entre empresas que se dedican a una actividad distinta en el desarrollo de productos o servicios en los que se aporta el valor añadido.

Alguna empresa resaltó los beneficios que puede tener para el sector conseguir una colaboración estructural entre los agentes:

"Creo que es una muestra de vinculación efectiva que no es oportunista, funcionamos mejor cuando cooperamos con otros agentes de forma sistemática, con un plan, con una misión común. Yo creo que es uno de los elementos claves a la hora de hacer más efectiva esta cooperación, pasar de lo oportunista a relaciones más estructurales, tanto con centros de investigación como con pequeñas empresas, que nos faciliten llevar la innovación con enfoque comercial a todos los clientes".

La cadena de valor del sector se divide en dos fases: una fase preindustrial donde se producen las infraestructuras lingüísticas a partir de la transformación de recursos básicos para las tecnologías del lenguaje. Estas infraestructuras son la base para la producción o el desarrollo de herramientas básicas de tecnologías del lenguaje.

La fase industrial comienza con el uso de esas herramientas o componentes básicos en el desarrollo de soluciones de tecnologías del lenguaje listas para comercializar.

El modelo de ingresos de los agentes del sector se basa fundamentalmente en la venta de servicios profesionales para el cliente final por trabajo finalizado, el acceso a subvenciones de I+D+i y en licencias asociadas a uso/acceso/trabajo realizado.

La modalidad de transferencia de software de código abierto es más utilizada por los agentes del sector para dirigirla a la sociedad o a los ciudadanos como clientes indirectos, a los centros de investigación y la universidad pública.

El producto o servicio acabado y los componentes de productos y servicios van dirigidos en mayor medida a empresas nacionales e internacionales. Y el asesoramiento y la formación en tecnologías del lenguaje a empresas nacionales y a la universidad pública.

Respecto al modelo de producción, la mayoría de los agentes consultados mantiene equipos para procesos de información.

En general, los agentes del sector no subcontratan servicios vinculados a las tecnologías del lenguaje a terceros (55%).

El 61,8% de las empresas del sector tienen departamento de I+D+i. Las empresas consultadas dedican un 6% del volumen de su facturación a inversión en I+D+i. Si atendemos al volumen de inversión que se corresponde con actividades del sector de tecnologías del lenguaje, representa el 10,1% del volumen de su facturación que se corresponde con la actividad.

Las redes de conocimiento en las que más participan los agentes del sector son las asociaciones especializadas (40,7%) y los clústers (35,6%).

El 60% de las empresas consultadas ha participado en líneas de investigación o desarrollo de negocio y el 84% de los centros de investigación ha realizado publicaciones científicas.

4 Tendencias y barreras del sector

4.1 Barreras del sector

En el informe sobre el sector de las tecnologías del lenguaje de la Agenda Digital para España[3] se afirmaba que "las consultoras internacionales pronostican un gran crecimiento del mercado mundial de aquí a 2020", no obstante, existen una serie de obstáculos que pueden dificultar el desarrollo del sector de tecnologías español.

En las entrevistas en profundidad realizadas y en los grupos de discusión celebrados se identificaron tres barreras principales a las que se enfrenta el sector: la falta de datos para entrenar los sistemas que alimentan las tecnologías del lenguaje, la falta de formación de profesionales del sector y el pequeño tamaño de las empresas que conforman el sector.

a) Falta de corpus

Las empresas y los centros de investigación apuntaron hacia la falta de datos o corpus para entrenar los sistemas de tecnologías del lenguaje como principal barrera a la que se enfrenta el desarrollo del sector en España. El desarrollo de aplicaciones de tecnologías del lenguaje depende de las infraestructuras lingüísticas de las que dispongan las empresas para alimentar y entrenar sus sistemas. Las empresas y los centros de investigación del sector señalan que existe una falta de datos a los que, generalmente, pueden acceder las grandes empresas ya que cuentan con los recursos económicos necesarios:

"El sector se está desarrollando enormemente y el mercado está aumentando, pero los que estamos aquí no nos lo comemos, una gran multinacional ha presentado 13 microservicios lingüísticos que permiten construir soluciones muy potentes. Entonces, yo creo que si queremos aspirar a que este sector sea mayor tenemos que expandirnos al mercado exterior. Hay desafíos muy grandes: tecnológicos, de facilidad de acceso al mercado, integración de las soluciones, etc. Nosotros construimos una parte, pero las grandes empresas que intervienen integran todo, y van a capturar todo

[3] http://www.agendadigital.gob.es/tecnologias-lenguaje/Bibliotecaimpulsotecnologiaslenguaje/Material%20complementario/Informe-Tecnologias-Lenguaje-Espana.pdf

el corpus y cuanto más corpus capturen mejor funcionarán sus sistemas. Entonces: datos, cómo accedemos más a los datos y cómo accedemos a una oferta".

Las empresas y los centros de investigación consultados expresaron que la Administración puede tener un papel facilitador en el problema de falta de datos, abriendo los datos y la información de la que dispone y poniéndolo a disposición pública.

Aspectos legales

El primer problema al que se enfrenta la apertura de datos por parte de la administración es el Reglamento General de Protección de Datos (RGPD) que entró en vigor recientemente y que concierne al tratamiento de los datos personales, las libertades públicas y los derechos fundamentales de las personas físicas, con la finalidad de preservar el honor, intimidad personal y familiar y el pleno ejercicio de los derechos personales. Todo esto es aplicable a los datos de carácter personal registrados en cualquier tipo de soporte físico susceptible de ser tratado, ya sea manual o informático. En este sentido, existe determinada información de la administración más complicada de tratar y de abrir, fundamentalmente la relacionada con la administración sanitaria o la justicia.

Interoperabilidad y formatos

Otro aspecto a tener en cuenta en la apertura de los datos por parte de la administración es el formato que se utilice y la plataforma desde la que se pueda acceder a ellos, es decir, existen un conjunto de decisiones de forma que envuelven a la infraestructura de datos abiertos que han de ser consideradas.

"Respecto al open data no se trataría tan solo de abrir los datos, sino que lo complicado está en preparar los procesos de acceso a esa información, es decir, el formato en el que se van a ofrecer, el repositorio desde el que se van a poner a disposición, el procedimiento de utilización de los datos, las aplicaciones que pueden tener, etc".

b) Falta de profesionales formados

La segunda barrera más comentada por las empresas y los centros de investigación consultados fue la de falta de formación de profesionales del sector. Esta falta de formación está vinculada al carácter multidisciplinario, asociado a la actividad de tecnologías del lenguaje, que precisa de perfiles mixtos formados por lingüistas y técnicos. Tal y como se señaló en el apartado de personal ocupado, existen dificultades para encontrar empleados formados en el área de tecnologías del lenguaje y las universidades no ofrecen titulaciones que combinen ambas disciplinas.

c) Pequeño tamaño de las empresas del sector

Para terminar, las empresas y los centros de investigación del sector expresaron que existe un obstáculo en su desarrollo relacionado con su tamaño. El sector de tecnologías del lenguaje está conformado por empresas y centros de investigación pequeños, lo que dificulta ganar cuota de mercado frente a competidores de otros países. Algunas empresas apuntaron que a nivel internacional existe una suerte de monopolio formado por grandes empresas que abarcan gran parte del mercado.

En este sentido, se torna sustancial la colaboración entre empresas y centros de investigación en la formación de redes colaborativas que faciliten un entorno innovador, donde puedan desarrollar aplicaciones competitivas. Además, la transversalidad de la aplicación de la actividad del sector de tecnologías del lenguaje a la mayoría de los sectores productivos podría correr el riesgo de convertir este tipo de tecnología en una *commodity*:

"Yo creo que al ser un sector que no es en sí finalista, sino que interviene de forma transversal en muchos sectores, muchas veces no tiene la visibilidad necesaria. Son tecnologías que están interviniendo en muchos procesos, pero a la vez quedan ocultas, y esto hace que las empresas que nos dedicamos al sector tengamos dificultad de visibilidad".

4.2 Tendencias y principales oportunidades del sector

En el trabajo cualitativo realizado a través de las entrevistas en profundidad y los grupos de discusión, se identificaron los sectores de actividad emergentes en la demanda de soluciones del sector, así como las oportunidades del sector.

En primer lugar, la oportunidad se encuentra en todas las actividades económicas y administrativas que puedan integrar las nuevas tecnologías, especialmente en aquellos sectores que requieran una interacción con el usuario final, dado que los usuarios demandan cada vez más una interacción más natural e inmediata.

A este respecto, destacan las tecnologías de procesamiento del habla, sistemas conversacionales y *chatbots* que permiten a las empresas que las incorporan ofrecer un servicio de conversación de texto con los usuarios, lo que conlleva un aumento de su disponibilidad y, a su vez, un ahorro de costes relacionado con *call centers* de atención al cliente.

En esta línea, cabría destacar el **sector de turismo**, intensivo en procesos de cara al público a través de diferentes canales (cara a cara, por teléfono, medios electrónicos, etc.), por lo que el lenguaje como interfaz es crítico para las actividades de este sector.

Por otra parte, destaca el **sector bancario** debido, por un lado, a la cantidad de documentos que tiene digitalizados (hipotecas, normativas internas…) que les interesa analizar, y por otro, motivados por su sección de atención al cliente, tanto por manejar los call centers, como por analizar los motivos por los que se pone el cliente en contacto con el banco, si expresa quejas, etc. Por tanto, la biometría aplicada al sistema financiero es un sector con gran potencial, ya que los bancos están demandando las tecnologías del lenguaje para mejorar sus formas de interaccionar con los clientes.

En segundo lugar, destaca el sector Big Data en una economía mundial que se encuentra en un proceso de transformación digital marcado por la necesidad de gestión de la ingente cantidad de datos y contenidos a través de Internet, redes sociales y medios digitalizados.

En este sentido, el procesamiento de esa información y su puesta en valor depende en gran medida del análisis del lenguaje natural. Existe una oportunidad estratégica en la aplicación de las tecnologías del lenguaje natural a los procesos de lenguaje no estructurado, una minería de datos de valor.

En tercer lugar, destaca el sector sanitario, concretamente la investigación biomédica, donde las tecnologías del lenguaje tienen tres aplicaciones detectadas como potenciales: el soporte a la decisión clínica, el enriquecimiento de colecciones de datos para la investigación y la codificación de diagnósticos.

En cuarto lugar, a través del análisis de textos, el análisis de sentimiento y la evaluación de perfiles se pueden extraer líneas de opinión en relación a un servicio o un producto, lo que se puede aplicar en diferentes actividades, como la evaluación de currículums, la evaluación de un sistema de consejos de inversión en bolsa, la predicción de resultados de elecciones, etc.

Por último, la generación automática de textos permite generar automáticamente informes de eventos deportivos, reseñas periodísticas a partir de datos, informes anuales de facturación de la empresa, etc.

4.3 El rol de la administración

Las líneas de financiación de I+D+i públicas

En general, las empresas y centros de investigación consultados en las entrevistas en profundidad y los grupos de discusión, conocen los programas y ayudas que ofrecen las administraciones públicas relacionados con el I+D+i, no obstante, consideran que existen factores que inhiben la participación de empresas y centros de investigación en subvenciones y programas públicos, algunos de ellos son factores transversales a cualquier sector de actividad. A continuación, se describen tres factores horizontales que frenan el acceso a programas de I+D+i en cualquier sector de actividad, que señalaron las empresas y los centros de investigación.

- El pequeño tamaño de las empresas, que implican escasos recursos económicos y de personal para afrontar la carga burocrática que lleva asociada la presentación de un proyecto.
  En este sentido, señalaron que para la micropyme y la pyme la relación entre la obtención de la subvención y el esfuerzo dedicado no compensa. Algunas empresas y centros apuntaron que, además, los ratios o índices de obtención de ayudas son muy bajos, por lo que directamente acuden a la financiación privada.

- Por otro lado, manifestaron el problema de empresas profesionales del proyectismo, para las que la subvención pública forma parte de su estructura de ingresos. Las empresas y los centros de investigación del sector de tecnologías del lenguaje afirmaron que sólo se pueden presentar a subvenciones que estén relacionadas con sus líneas de investigación o líneas de trabajo, no tienen los recursos necesarios para desviarse de su ámbito de actuación. Por ello, señalaron que una particularidad que ocurre en el sector es que en las convocatorias no se encuentran referencias explicitas a las tecnologías del lenguaje.

- En último lugar, las empresas y los centros de investigación manifestaron la falta de evaluación posterior de las subvenciones que se aprueban, lo que impide medir el impacto real que tienen en las empresas o los centros de investigación beneficiarias.
  "Nosotros construimos una plataforma que integraba tecnologías lingüísticas para traducción con tecnologías de administración, con herramientas como una framework de ERP de gestión, de control de presupuestación. Eso no está en las convocatorias, esas estructuras reales que nos hacen falta no están. Bueno pues nos financiaron para eso, para montar una plataforma, bueno pues esa tecnología, esa gestión y ese crédito cuando en aquel momento éramos 18

personas, ha dado 40 puestos de trabajo 10 años después, pero eso no hay constancia en ningún sitio".

Otro factor que identificaron los centros de investigación y las empresas y que podría estar más relacionado, si no con el sector de las tecnologías del lenguaje, con el sector TIC, es que las convocatorias están orientadas al desarrollo de bienes tangibles, no se tiene en cuenta las características y particularidades que tiene el desarrollo de software con respecto a la industria tangible, que es totalmente medible, cuantificable y comparable.

"En muchos de esos pliegos estamos encontrando que el peso es más el precio que lo que es la parte técnica, o sea parece que está más pensado esto para hacer una carretera que para hacer un desarrollo de software, entonces como parece que estamos heredando digamos algún tipo de proceso de hacer una construcción de algo material, físico, una carretera, un puente, una casa, donde tú de alguna manera estás poniendo una serie tabulada de materiales que puedes comparar, pero el desarrollo de software no se mide así, es mucho más intangible".

Una de las particularidades está relacionada con la dificultad de aplicar un criterio de calidad a un desarrollo de software tan subjetivo, si bien es más sencillo medir la funcionalidad.

Otra cuestión que señalaron las empresas y los centros de investigación relacionada con las convocatorias de subvenciones públicas, es que se demanda la solución o el producto de tecnologías del lenguaje entero, es decir, desde la creación de las infraestructuras lingüísticas hasta el desarrollo de la aplicación.

Para una empresa basada en el conocimiento o el *know how*, entregar sus infraestructuras lingüísticas o lo que es lo mismo, su propiedad intelectual donde ha estado trabajando durante años por un único servicio, no le resulta rentable. Más cuando este servicio se ofrece a precio de aplicación, sin tener en cuenta el valor económico que tienen las infraestructuras lingüísticas. En este sentido, señalaron que las convocatorias deberían reorientarse hacia componentes tecnológicos atomizados que permitan ofrecer un servicio o producto muy específico.

Demanda de tecnologías del lenguaje de la Administración

Con el objetivo de conocer la demanda de tecnologías del lenguaje por parte de la Administración se entrevistó al responsable del Plan de Impulso de las Tecnologías del Lenguaje (Plan TL) de la Secretaria de Estado para la Sociedad de la Información y la Agenda Digital (SESIAD).

Las medidas propuestas en el Plan TL se organizan en cuatro ejes, entre los que se encuentra el Eje IV: Proyectos Faro. Los proyectos faro son proyectos emprendidos por las Administraciones Públicas de aplicación de las tecnologías del lenguaje en sectores estratégicos que pretenden servir de demostración de sus capacidades y beneficios, generar industria y generar recursos reutilizables en otros proyectos. Estos proyectos impulsados por la SESIAD están dirigidos a los llamados sectores verticales.

Estos proyectos se instrumentalizarán a través de la creación de una oficina técnica general para cada uno de los sectores verticales. Las oficinas técnicas pretenden reunir a grupos de expertos o competentes que asesoren a cada administración en la creación de productos piloto que después puedan lanzarse al mercado en forma de compra pública innovadora, como una suerte de catalizador.

Los componentes de esos productos piloto han de ser totalmente interoperables, atómicos y modulares, de manera que se puedan dividir cada uno de los componentes del producto para su comercialización.

La necesidad de atomizar los productos viene motivada por la última fase de intervención de la oficina técnica general, la evaluación de dichos productos sobre una base científica. Esta evaluación permitirá que la administración compruebe los componentes de producto que han funcionado y los que se podría mejorar. Cabe señalar, que los expertos o competentes que participen de la oficina técnica no podrán ser en modo alguno receptores de los productos cuando salgan al mercado, para evitar la distorsión que podría conllevar.

Las oficinas técnicas generales se crearán en cada uno de los sectores verticales. A continuación, se describen los contactos que se han establecido entre la SESIAD y las administraciones en los sectores verticales, y la información que han aportado los representantes de dichos sectores en las entrevistas en profundidad realizadas.

Justicia

Respecto a la administración de justicia se han establecido contactos con el Consejo General del Poder Judicial, concretamente con el Centro de Documentación Judicial (CENDOJ), donde se encuentran digitalizadas hasta 6 millones de sentencias del Tribunal Supremo, la Audiencia Nacional, el Tribunal Superior de Justicia, la Audiencia Provincial y Tribunales Militares y Unipersonales, y con el propio Ministerio de Justicia. Por otro lado, se han establecido contactos para que la oficina técnica general esté formada por expertos de la Universidad del País Vasco.

Se entrevistó a un representante del Ministerio de Justicia para conocer su perspectiva acerca de la implementación de las tecnologías del lenguaje en la administración de justicia.

Desde su perspectiva, las tecnologías del lenguaje son importantes en la gestión y la transcripción de la administración de justicia, no obstante, el lugar que ocupan las tecnologías del lenguaje en el proceso de digitalización de la administración de justicia es todavía experimental. Fundamentalmente, porque la administración de justicia se encuentra en una etapa de digitalización que, una vez conseguida, llevará consigo la incorporación de la inteligencia artificial a los procesos de digitalización, que es el marco de la incorporación de las tecnologías del lenguaje.

La forma en la que funciona la administración de justicia en la consecución de la transformación digital es a través de dos vías: por un lado, el Ministerio de Justicia y las Comunidades Autónomas son las que ponen a disposición los medios electrónicos, por lo que se torna sustancial generar un marco común con las comunidades autónomas de manera que se pueda conseguir una digitalización más armonizada y en el mismo camino hacia la transformación digital; y, por otro lado, el Centro de Documentación Judicial (CENDOJ) que cuenta con una gran base de datos de sentencias digitalizadas, por lo que se trata del activo más importante que tiene la administración de justicia para aplicar las tecnologías del lenguaje, precisamente, donde se han iniciado los contactos para el desarrollo de los proyectos faro.

Sanidad

Respecto a la administración sanitaria se han establecido contactos con el Servicio Andaluz de Salud, con la Comunidad Autónoma de Baleares, y con el Centro Nacional de Investigaciones Oncológicas (CNIO) y el Barcelona Supercomputer Center – Centro Nacional de Supercomputación (BSC-CNS) para ofrecer expertos que trabajen en la oficina técnica general.

En las entrevistas en profundidad se contactó con un representante del Servicio Andaluz de Salud para conocer su perspectiva acerca de la implementación de las tecnologías del lenguaje en la administración sanitaria.

El Servicio Andaluz de Salud está adscrito a la Consejería de Salud de la Junta de Andalucía y cuenta entre sus líneas de investigación con la informática médica y el servicio de lenguaje natural en informática médica, por lo que han desarrollado proyectos de investigación e innovación en esta área.

La administración sanitaria tiene una gran fuente de información relacionada con los registros clínicos o registros de salud, es decir, toda la información que se recoge en el contexto de la asistencia sanitaria. Gran parte de esa información está digitalizada, por lo que su reutilización representa una oportunidad

para la investigación y la propia asistencia sanitaria, esto es, el acceso a la información, su recuperación, su normalización y su computación para usos avanzados en asistencia sanitaria e investigación biomédica.

Para la investigación biomédica y el uso en asistencia sanitaria de la información y los datos que pertenecen a la administración sanitaria se utilizan tanto técnicas de traducción automática como procesamiento del habla y sistemas conversacionales y procesamiento del lenguaje natural. Concretamente, la administración sanitaria requiere de las herramientas de traducción automática en tanto se atiende a personas de diferentes nacionalidades y con diferentes idiomas. Por otra parte, la administración está demandando el desarrollo de asistentes virtuales en el área de procesamiento del habla para realizar cuestionarios de seguimiento a los pacientes o para evitar una visita hospitalaria a los centros sanitarios. Para terminar, el principal recurso con el que cuenta la investigación biomédica es el acceso a los recursos médicos a través del procesamiento del lenguaje natural.

Respecto a los proyectos específicos en los que se está trabajando desde el Servicio Andaluz de Salud relacionados con las tecnologías del lenguaje, se podrían distinguir tres usos diferenciados:

- En primer lugar, como soporte a la decisión clínica se utiliza la información de los textos clínicos, se normaliza y se computa en base a herramientas justificadas que aplican reglas de decisión en base a guías clínicas.

  Para aplicar las reglas de decisión se tienen en cuenta diversos factores probabilísticos pero para realizar su computación, la información de entrada ha de entenderse, y es ahí donde entran las tecnologías del lenguaje, ya que esa información que se encuentra en los textos está condicionada por los sinónimos locales, el uso de expresiones poco regulares, el contexto de negación o el contexto de especulación, para las que es necesario aplicar técnicas de procesamiento del lenguaje natural que permitan identificar, clasificar y normalizar la información de manera que sirva de entrada a las máquinas de decisión y hagan recomendaciones adecuadas a los médicos en esos sistemas de soporte a la decisión clínica.

- Por otro lado, **la** información clínica y sanitaria sirve de enriquecimiento de colecciones de datos para la investigación clínica, de esta manera que se utiliza una colección de datos de una cohorte de pacientes que cumplan ciertos criterios, y se recogen los datos que permitan encontrar patrones de relación a través del procesamiento del lenguaje natural.

- Por último, las tecnologías del lenguaje aplicadas a la investigación biomédica permiten codificar información relevante para la gestión de los hospitales, a través de la codificación de

diagnósticos y procedimientos en base a unos códigos internacionales normalizados[4] y la clasificación internacional de enfermedades.

A este respecto, en el sector sanitario existen unos estándares de normalización de terminología a nivel mundial que permiten que la aplicación de las tecnologías del lenguaje a la investigación biomédica se traduzca en soluciones competitivas.

Por otra parte, señaló el problema que tiene la generación de corpus en el ámbito sanitario, que precisa de mucha especialización en dominios concretos, es decir, enfermedades y especialidades concretas. En este sentido, apuntó una barrera importante para la generación de corpus y es el marco legal, que especialmente afecta a este sector donde la información que es vulnerable de verse afectada por el Reglamento General de Protección de Datos (RGPD) y, además, tiene un marco legal específico condicionado por la Ley General de Sanidad o la Ley de Investigación biomédica. Además, a esa situación se suma que las comunidades autónomas tienen competencias en sanidad, por lo que no existe un marco regulatorio homogéneo a nivel estatal que permita la colaboración en la creación de corpus.

Inteligencia competitiva

La inteligencia competitiva sectorial pretende, por un lado, evaluar los proyectos de I+D+i que se presenten a la administración, por otro lado, ayudar al decisor político en las decisiones relacionadas con hacia dónde dirigir la inversión de I+D+i y, por último, evaluar el impacto de las ayudas que se aprueben.

- La SESIAD está trabajando en una herramienta de tecnologías del lenguaje que permite al evaluador de proyectos I+D+i clasificar los proyectos que se presenten por temáticas, a través del procesamiento del lenguaje natural, con el objetivo de que pueda realizar clasificaciones, comparaciones y detectar el posible fraude en la presentación de los mismos proyectos por empresas en distintas ventanas.
- La Secretaria de Estado de Investigación Desarrollo e Innovación (SEIDI) está trabajando en una herramienta de administración clúster que sirva de apoyo al decisor político a la hora de

[4] CIE-9-MC: es la traducción oficial de ICD-9-MC International Classification of Diseases, creada para facilitar la codificación de morbimortalidad en los hospitales. CIE-9-MC es un acrónimo de Clasificación Internacional de Enfermedades, novena revisión, modificación clínica. Se trata de una clasificación de enfermedades y procedimientos utilizada en la codificación de información clínica derivada de la asistencia sanitaria, principalmente en el entorno de hospitales y centros de atención médica especializada.

decidir en qué ámbitos invertir el dinero público. A través de una clasificación de las temáticas de los pliegos que se publican en España, en Europa y en Estados Unidos, el decisor político puede observar para cada temática de actuación, su evolución en el tiempo, tanto a nivel europeo como americano, y así poder hacerse una idea de las líneas en las que están invirtiendo en EEUU, o de los ámbitos en los que se ha invertido menos en España en los últimos años. En general, esta herramienta permite conocer la colaboración entre organizaciones por temáticas, los proyectos europeos que se están desarrollando y cómo evolucionan cada una de las temáticas de investigación.

- Por último, la evaluación del impacto de las ayudas no tiene un enfoque económico entendido como dinero generado por dinero invertido, sino que lo que se pretende es que se pueda comprobar y así evitar el solapamiento entre organismos en los mismos ámbitos de actuación.

Cultura

El ámbito de la cultura ha sido la última incorporación a los sectores verticales de actuación, la SESIAD ha firmado un convenio de colaboración con la Real Academia Española (RAE) y mantiene contactos con la Biblioteca Nacional.

En las entrevistas en profundidad se contactó con un representante del Centro de estudios de la Real Academia Española para conocer su perspectiva acerca de la implementación de las tecnologías del lenguaje en la Real Academia Española.

El Centro de estudios cuenta con un departamento de lingüística computacional en el que se están desarrollando diversas actividades relacionadas con las tecnologías del lenguaje desde finales de los años 90 entre las que se encuentra, el análisis lingüístico de corpus, la generación de tecnología lingüística o la construcción de noticias periodísticas y análisis lingüístico.

Para el desarrollo de estas actividades se aplican principalmente tecnologías de procesamiento del lenguaje natural. Además, el departamento de lingüística computacional trabaja en la creación de recursos para las tecnologías del lenguaje en dos sentidos, por un lado, en la creación de corpus puestos a disposición del público para su consulta y, por otro lado, en la creació**n de recursos internos**, como, por ejemplo, lexicones de análisis morfosintácticos, que utilizan junto a la tecnología que producen internamente para acotar esos corpus que se ponen a disposición del público.

En la actualidad, el departamento de lingüística computacional está trabajando principalmente en dos proyectos:

- A nivel europeo, están trabajando en un proyecto llamado Ilexis, en el que colaboran hasta diecisiete participantes, entre los que se encuentran academias, instituciones relacionadas con lenguas internacionales y grupos de investigación. A través del análisis de datos enlazados y el procesamiento del lenguaje, este proyecto pretende conectar distintos diccionarios de varias lenguas a través de un recurso lingüístico.
- A nivel nacional, el centro de estudios de la RAE ha lanzado, recientemente, una plataforma bajo suscripción llamada Enclave.rae.es, en la que se pone a disposición de los usuarios corpus avanzado y herramientas de base lingüística. Concretamente, desde el departamento de lingüística computacional han incorporado un verificador ortográfico y de estilo, y están trabajando en una consulta avanzada del diccionario y en un módulo de verificación gramatical.

Turismo

Para finalizar con el análisis de la demanda de tecnologías del lenguaje de la administración se contactó con un representante de la Sociedad Mercantil Estatal para la Gestión de la Innovación y las Tecnologías Turísticas (SEGITTUR) dependiente del Ministerio de Industria Turismo y Comercio, y adscrita a la Secretaría de Estado de Turismo, responsable de impulsar la innovación (I+D+i) en el sector turístico español.

Desde su perspectiva, el sector de turismo es un sector intensivo en personas, es decir, en procesos cara al público a través de diferentes canales (cara a cara, por teléfono, medios electrónicos, etc.), por lo que el lenguaje como interfaz es crítico para el sector turismo en toda la etapa del viaje, el antes, el durante y el después.

La propia naturaleza del sector en el que operadores privados y operadores intermedios participan de toda la etapa del viaje, desde los hoteles, las agencias de viaje, las empresas de transporte o la industria de restauración, hace que el lenguaje sea un factor crítico en el desarrollo del sector.

Además, la información que se maneja en el sector turístico es más fácilmente adaptable al uso de las tecnologías del lenguaje por tres motivos:

- Por un lado, se trata de compras de servicios sobre bases muy estandarizadas.
- Por otro lado, es un sector que está muy digitalizado, por lo que resulta más sencillo que en otros sectores (como justicia) introducir la capa de este tipo de tecnologías en los grandes intermediarios de e-commerce que conforman el sector. La tecnificación del sector permite

que las tecnologías del lenguaje se puedan aplicar a muchas capas: grandes intermediarios, industria hotelera, industria de transportes (grandes aerolíneas), agencias de viaje e industria de restauración.

- Y, por último, **el** sector turístico no maneja información con datos personales por lo que no le afectaría de igual forma que a otras administraciones el marco regulatorio de los datos abiertos.

Desde SEGITTUR se ha propuesto la creación de una plataforma digital de tecnologías del lenguaje aplicadas al sector turístico, que tiene como objetivo ofrecer recursos lingüísticos de valor añadido para mejorar la investigación, desarrollo e innovación de soluciones tecnológicas a través de motores de reconocimiento automático, motores de conversión texto a voz de calidad, motores de traducción automática para dominios semánticos restringidos, gramáticas estadísticas para la mejora de los procesos de reconocimiento, generación de nuevos módulos para la extracción de tópicos, generación de resúmenes, etc.

La plataforma pretende ser multimodal (voz, video, texto), multilenguaje (varios idiomas), multiservicio (diferentes servicios en la nube que hagan uso de los recursos necesarios, bases de datos, open data, big data, internet IoT, etc), multidispositivo (tablets, smartphones, PC), ofreciendo las siguientes prestaciones:

- Asistente turístico virtual: interacción con la plataforma en lenguaje natural.
- Traducción simultánea: tanto texto-texto, como voz-voz, en el entorno semántico del sector turismo.
- Web semántica: que posibilita la publicación y búsqueda inteligente de información turística.
- Información pública, open data / big data: facilitando el acceso a la información turística pública.
- Canales de acceso: web, contact center, dispositivos móviles, redes sociales.

Por tanto, se combinaría la tecnología, la ingeniería lingüística y el conocimiento adquirido del sector de turismo a partir de los recursos multilingües que pertenecen a la Secretaría de Estado de Turismo y TURESPAÑA.

Barreras

Las principales barreras del sector identificadas están vinculadas a la falta de datos o corpus para entrenar los sistemas de tecnologías del lenguaje, los aspectos legales de la apertura de datos, la interoperabilidad y el formato de los datos abiertos, la falta de profesionales formados debido a la multidisciplinariedad de la actividad TL y el pequeño tamaño de las empresas que conforman el sector.

Oportunidades

Las principales oportunidades del sector se encuentran, en primer lugar, en aquellos sectores de actividad que requieran una interacción con el usuario final, especialmente, el sector turismo y bancario.

En segundo lugar, existe una oportunidad estratégica en la aplicación de las tecnologías del lenguaje natural a los procesos de lenguaje no estructurado, una minería de datos de valor. Así lo demuestra que el 60% de los agentes consultados mantiene una línea de investigación vinculada a la minería de datos (análisis de sentimiento).

En tercer lugar, destaca el sector sanitario y la aplicación de las tecnologías del lenguaje a la investigación biomédica.

Factores que inhiben la participación de los agentes en líneas de financiación pública:

En primer lugar, los ratios o índices de obtención de ayudas son muy bajos para empresas de pequeño tamaño y con escasos recursos para afrontar la carga burocrática que implican estos proyectos.

En segundo lugar, los agentes expresaron que los pliegos están orientados a bienes tangibles, no se tiene en cuenta las características y particularidades que tiene el desarrollo de software con respecto a la industria tangible, que es totalmente medible, cuantificable y comparable.

Factores que inhiben la participación en líneas de financiación de I+D+i públicas:

El esfuerzo dedicado para la obtención de una subvención pública no compensa a las microempresas y pymes del sector.

Las convocatorias no están orientadas explícitamente a la actividad del sector de tecnologías del lenguaje y las empresas suelen presentarse a convocatorias que están relacionadas con sus líneas de investigación o actividad.

Falta de seguimiento de los proyectos que resultan subvencionados, lo que impide medir el impacto real que tienen en las empresas beneficiarias.

Las convocatorias están orientadas al desarrollo de bienes tangibles.

Las convocatorias deberían reorientarse hacia componentes tecnológicas atomizadas que permiten a las empresas ofrecer sus servicios sin entregar todo su know how.

5 Análisis DAFO

El análisis DAFO (Debilidades, Amenazas, Fortalezas, Oportunidades) es una herramienta utilizada para realizar un diagnóstico de la situación de un determinado proyecto, empresa, institución, etc., donde se analiza la situación interna (Debilidades y Fortalezas) y la situación externa (Amenazas y Oportunidades), para poder actuar de cara al futuro y seguir mejorando. Para realizar este análisis se confecciona una matriz donde se recoge la situación actual del sector de las tecnologías del lenguaje con respecto a cada uno de estos elementos.

Cabe señalar que en la publicación del Plan TL (2015) se presentó un análisis DAFO exhaustivo del sector de tecnologías del lenguaje[5].

A continuación, se describe el análisis DAFO realizado en el presente estudio, a partir de la información recogida a través de las diferentes técnicas cualitativas aplicadas en el estudio, las entrevistas en profundidad y los grupos de discusión, y el análisis DAFO que se realizó en 2015 por la SESIAD.

En primer lugar, **se** han identificado las fortalezas internas del sector de tecnologías del lenguaje español.

- España es un país puntero en las tecnologías del lenguaje, al tratarse de una sociedad plurilingüe.
- En comunidades autónomas como Galicia, País Vasco, Cataluña, Comunidad Valenciana, existen grupos de investigación con experiencia en procesamiento del lenguaje natural donde se concentra parte de la capacidad del sector.

Estas dos fortalezas que se han identificado en el presente estudio coinciden con las identificadas por la SESIAD en 2015:

- Desarrollo de líneas de investigación en procesamiento de lenguaje natural en España que abarcan casi todos los ámbitos en los que se trabaja actualmente a nivel internacional. Existen recursos y herramientas propias, consolidadas y robustas para hacer el procesamiento básico de lenguaje natural y traducción automática para el castellano, catalán, vasco y gallego, además del inglés y se dispone de amplia información del sector público susceptible de convertirse en recursos lingüísticos.

[5] http://www.agendadigital.gob.es/tecnologias-lenguaje/BibliotecaimpulsotecnologiaslenguaJe/Detalle%20del%20Plan/Plan-Impulso-Tecnologias-Lenguaje.pdf

- Disponibilidad de investigadores españoles que participan en proyectos, asociaciones y grupos de estandarización europeos e internacionales. Existen más de 30 grupos de investigación consolidados y organizados en asociaciones y redes, que han sido el origen de 9 spin-off que abarcan casi todos los ámbitos en los que se trabaja actualmente a nivel internacional.
- Gran experiencia en gestión de multilingüismo de las empresas españolas y las Administraciones Públicas, lo que puede ser un modelo exportable.

Por otra parte, se han identificado cinco fortalezas más del sector de tecnologías del lenguaje que no se identificaron en el Plan de Impulso de 2015:

- La mayoría de las empresas y los centros de investigación del sector llevan entre 11 y 20 años dedicándose a la actividad de tecnologías del lenguaje, lo que muestra una actividad antigua y consolidada.
- Los agentes del sector han contratado personal especializado durante 2017, lo que muestra que el negocio o la actividad TL está aumentando o adquiriendo mayor peso en las líneas de actividad y/o investigación de los agentes.
- Desde la perspectiva de las ventas, se ha detectado un sector en auge en la medida en que poco más de la mitad de los agentes aumentaron su volumen de clientes en 2017.
- El volumen de facturación de la actividad de tecnologías del lenguaje en 2016 fue de aproximadamente 200 millones de euros.

La Red Temática en Tecnologías del Habla[6] se creó en el año 2001 con el propósito de agrupar a todos los agentes en el ámbito español relacionados con las Tecnologías del Habla. Con los objetivos de facilitar el intercambio y transferencia de conocimientos, la cooperación entre los agentes y la coordinación entre las infraestructuras y la difusión de la red, se creó un libro blanco en el que figuran los medios humanos y materiales con que cuenta la red.

A continuación, se ofrece una tabla resumen con las fortalezas identificadas en 2015 y las identificadas en el presente estudio.

[6] http://www.rthabla.es/

Fortalezas Plan Impulso 2015	Fortalezas Estudio TL 2018
• Desarrollo de líneas de investigación en procesamiento de lenguaje natural en España que abarcan casi todos los ámbitos en los que se trabaja actualmente a nivel internacional. Existen recursos y herramientas propias, consolidadas y robustas para hacer el procesamiento básico de lenguaje natural y traducción automática para el castellano, catalán, vasco y gallego, además del inglés y se dispone de amplia información del sector público susceptible de convertirse en recursos lingüísticos. • Disponibilidad de investigadores españoles que participan en proyectos, asociaciones y grupos de estandarización europeos e internacionales. Existen más de 30 grupos de investigación consolidados y organizados en asociaciones y redes, que han sido el origen de 9 spin-off que abarcan casi todos los ámbitos en los que se trabaja actualmente a nivel internacional. • Posibilidad de establecer modelos de colaboración con centros investigadores de forma rápida gracias a los programas nacionales de transferencia (CENIT, CIEN, Doctorados industriales). • Gran experiencia en gestión de multilingüismo de las empresas españolas y las Administraciones Públicas, lo que puede ser un modelo exportable. • Posición de liderazgo de España en el mercado potencial del español con cerca de 470 millones de habitantes, 54 en EE. UU., por su situación económica. Se han abierto oficinas en EE. UU. e Iberoamérica por parte de varias empresas españolas, demostrando que la internacionalización de la tecnología desarrollada en España es viable y que existe un sector TIC español potente que está demostrando su capacidad para competir globalmente. • España pertenece a las redes Iberoamericanas, con prestigio en la sociedad, de las instituciones relacionadas con las lenguas españolas (RAE, IEC, etc.) que colaboran internacionalmente en la labor de regulación del lenguaje. • Bajo coste de reutilizar y visibilizar todos los materiales existentes para los lenguajes de especialidad, dada la estructura y organización ya existente y demostrada eficacia de la experiencia en acciones estratégicas en los planes nacionales. • Existencia de bases de datos terminológicas, sinónimos, tesauros y topónimos, exportables globalmente. Existencia de un tejido industrial en	• España es un país puntero en las tecnologías del lenguaje, al tratarse de una sociedad plurilingüe. • En comunidades autónomas como Galicia, País Vasco, Cataluña o Comunidad Valenciana, existen grupos de investigación con experiencia en procesamiento del lenguaje natural donde se concentra parte de la capacidad de investigación del sector. • La mayoría de las empresas y los centros de investigación del sector llevan entre 11 y 20 años dedicándose a la actividad de tecnologías del lenguaje, lo que muestra una actividad antigua y consolidada. • Los agentes del sector han contratado personal especializado durante 2017, lo que muestra que el negocio o la actividad TL está aumentando o adquiriendo mayor peso en las líneas de actividad y/o investigación de los agentes. • Desde la perspectiva de las ventas, se ha detectado un sector en auge en la medida en que poco más de la mitad de los agentes aumentaron su volumen de clientes en 2017. • El volumen de facturación de la actividad de tecnologías del lenguaje en 2016 fue de aproximadamente 200 millones de euros. • La Red Temática en Tecnologías del Habla se creó en el año 2001 con el propósito de agrupar a todos los agentes en el ámbito español relacionados con las Tecnologías del Habla. Con los objetivos de facilitar el intercambio y transferencia de conocimientos, la cooperación entre los agentes y la coordinación entre las infraestructuras y la difusión de la red, se creó un libro blanco en el que figuran los medios humanos y materiales con que cuenta la red.

traducción automática con experiencia que tiene como clientes prioritarios las Administraciones Públicas. • Existencia de iniciativas de traducción automática ya en marcha, como la plataforma Plata. • Adhesión de España a la directiva europea de reutilización de datos de las Administraciones Públicas (RISP). • Experiencia en la participación en iniciativas europeas como Meta-Share, Clarin-Eric, ELRA, etc., y en proyectos europeos como OPENER, NEWSREADER, QT-Leap, etc.	

En segundo lugar, se han identificado nuevas debilidades internas del sector de tecnologías del lenguaje español respecto a las identificadas en 2015:

- Las empresas que se dedican al sector necesitan visibilidad ya que las actividades a las que se dedican las tecnologías del lenguaje no son finalistas, sino que intervienen de forma transversal en muchos procesos y sectores, por lo que pueden quedar ocultas y corren el riesgo de ser percibidas como una "commodity".
- Las empresas y centros de investigación necesitan generar corpus con los que entrenar sus sistemas.
- El marco regulatorio del uso de los datos públicos en España pone en riesgo la inversión de empresas en tecnologías del lenguaje y la investigación y desarrollo de soluciones por parte de los grupos de discusión.
- Creencia de que las tecnologías del lenguaje amenazan la privacidad de los ciudadanos y las administraciones.
- Existe un problema de crecimiento de las start-ups y microempresas en España.
- Los centros de investigación tienen problemas para crear spin-off: es un recorrido muy largo que consume el tiempo que tienen que dedicar a la investigación para sobrevivir, además, para crear una spin-off un centro de investigación necesita tener compromiso de compra.
- Se necesita inversión en formación de especialistas en tecnologías del lenguaje: son perfiles mixtos, con conocimientos en servicios lingüísticos y servicios técnicos, por su grado de especialización son difíciles de encontrar.
- Escasa inversión en I+D+i relacionada con las tecnologías del lenguaje.

- Las empresas y los centros de investigación no utilizan las líneas de subvención de proyectos I+D+i públicas.
- La lengua castellana tiene recursos al nivel de las demás lenguas hegemónicas pero la lengua inglesa es la predominante en cuanto a investigación y desarrollo.

Respecto a las debilidades detectadas en el Plan en 2015, coinciden en su mayoría con las identificadas en el presente estudio, exceptuando algunas debilidades como, por ejemplo:

- Carencia de una norma específica de interoperabilidad.
- Reducida definición de la estrategia de comercialización generando grandes dificultades para acceder al mercado.
- Falta de mayor cultura RISP en los diferentes colectivos.

Se entiende que estas debilidades se han mantenido, ya que no se ha contrastado lo contrario.

A continuación, se expone la tabla comparativa de las debilidades del sector detectadas en el presente estudio y las que se identificaron en 2015:

Debilidades Plan Impulso 2015	Debilidades Estudio TL 2018
• Insuficiente colaboración entre empresas y entre empresas y grupos de investigación impidiendo la reutilización de datos y herramientas y multiplicando la inversión que realizan las empresas, restándole efectividad en otros ámbitos como la promoción comercial. Falta de conocimiento e inversión coordinada y escasa compartición de herramientas de amplia cobertura entre las empresas y las academias, lo que dificulta la implantación de métodos que garanticen la reutilización, generando duplicidades y dispersión de esfuerzos en la construcción de corpus, herramientas, etc. • Déficit de interdisciplinariedad debido a la baja interacción entre lingüistas e informáticos en la creación y compartición de recursos y aplicaciones, motivado por la rigidez estructural universitaria y la falta de centros de investigación sobre procesamiento de lenguaje natural y traducción automática. • Discontinuidad en la financiación de la investigación en procesamiento de lenguaje natural y traducción automática, lo que dificulta el progreso en la	• Las empresas que se dedican al sector necesitan visibilidad ya que las actividades a las que se dedican las tecnologías del lenguaje no son finalistas, sino que intervienen de forma transversal en muchos procesos y sectores, por lo que pueden quedar ocultas y corren el riesgo de ser percibidas como una “commodity”. • Las empresas y centros de investigación necesitan generar corpus con los que entrenar sus sistemas. • El marco regulatorio del uso de los datos públicos en España pone en riesgo la inversión de empresas en tecnologías del lenguaje y la investigación y desarrollo de soluciones por parte de los grupos de discusión. • Creencia de que las tecnologías del lenguaje amenazan la privacidad de los ciudadanos y las administraciones.

investigación y el mantenimiento de equipos de trabajo especializados. • Insuficiente investigación básica y desarrollo tecnológico en torno al procesamiento de lenguaje natural o a la traducción automática por parte de las Agencias Públicas de Investigación y disminución del número de grupos de investigación en procesamiento de lenguaje natural y traducción automática en grandes empresas en España (IBM, TIC, etc.). • Reducida disponibilidad de recursos y herramientas para el español hispanoamericano. • Existencia de una debilidad en cuanto al tratamiento de textos especializados ya que el castellano no es un idioma mayoritario en literatura científica y patentes. • Escaso conocimiento de los estándares, licencias y modelos de negocio ya consensuados en Europa por las empresas en España. • Falta de reconocimiento de esta disciplina en la academia nacional pese a ser un ámbito de investigación interdisciplinar. La formación está incluida en diferentes especialidades, lo que dificulta su visibilidad. • Falta de un mayor conocimiento de las características de esta tecnología con poco coste de licencia de programas, pero alto coste de inversión en localización y adaptación (lengua y dominios de actividades) para lograr procesos mucho más eficientes en múltiples áreas de aplicación en la Administración. • Carencia de una norma específica de interoperabilidad. • Reducida definición de la estrategia de comercialización generando grandes dificultades para acceder al mercado. • Falta de mayor cultura RISP en los diferentes colectivos.	• Existe un problema de crecimiento de las start-ups y microempresas en España. • Los centros de investigación tienen problemas para crear spin-off: es un recorrido muy largo que consume el tiempo que tienen que dedicar a la investigación para sobrevivir, además, para crear una spin-off un centro de investigación necesita tener compromiso de compra. • Se necesita inversión en formación de especialistas en tecnologías del lenguaje: son perfiles mixtos, con conocimientos en servicios lingüísticos y servicios técnicos, por su grado de especialización son difíciles de encontrar. • Escasa inversión en I+D+i relacionada con las tecnologías del lenguaje. • Las empresas y los centros de investigación no utilizan las líneas de subvención de proyectos I+D+i públicas. • La lengua castellana tiene recursos al nivel de las demás lenguas hegemónicas pero la lengua inglesa es la predominante en cuanto a investigación y desarrollo.

En tercer lugar, se han identificado las oportunidades externas del sector de tecnologías del lenguaje español:

- El sector de tecnologías del lenguaje no tiene impedimentos para la exportación similares a otros sectores de actividad que comercializan con productos.
- En un mercado con grandes proveedores, como Google o Microsoft, es prácticamente imposible para una pequeña empresa española entrar a competir en el mercado inglés, por lo

que se torna esencial abrirse al mercado de la lengua castellana, esto es al mercado latinoamericano.

- El castellano es la lengua con mayor índice de aceleración en este tipo de tecnologías: es una lengua que está bien posicionada, hay recursos, hay bases de datos y hay herramientas. Es una lengua conocida que se utiliza, que tiene presencia.
- La situación de la lengua castellana a nivel internacional mejora respecto a su situación a nivel europeo gracias a la introducción del mercado latinoamericano, que la mayoría de los agentes coinciden en señalar como un nicho de mercado del sector
- Se está produciendo un auge de los sistemas neurales para los que se necesita asistencia computacional.

A continuación, se exponen la tabla comparativa las oportunidades del sector detectadas en el presente estudio y las que se identificaron en 2015:

Oportunidades Plan Impulso 2015	Oportunidades Estudio TL 2018
• Gran momento en Europa para el desarrollo del procesamiento de lenguaje natural y traducción automática con la última incorporación de nuevos países. Infraestructura, asociaciones, fundaciones y distribuidores han trabajado ya de forma colaborativa los aspectos formales: estándares, licencias y modelos de negocio. En la actualidad, existen modelos europeos e internacionales de oferta de datos lingüísticos en portales de datos abiertos a los que sumarse. • Demanda urgente de aplicaciones relacionadas con social media, big data y los datos abiertos, creando objetivos a corto plazo que ya pueden ser satisfechos en consorcios mixtos de desarrollo, lo que proporcionará gran visibilidad al área. • Disponibilidad de múltiples áreas de aplicación: turismo, sanidad, justicia, educación, etc., donde optimizar y sistematizar procesos horizontales que pueden servir de efecto demostrador y aprendizaje para proyectos futuros, por su posibilidad de generar recursos reutilizables. • Es posible generar valor mediante la definición de conjuntos de datos abiertos de interés lingüístico dentro de la estrategia RISP.	• El sector de tecnologías del lenguaje no tiene impedimentos para la exportación similares a otros sectores de actividad que comercializan con productos. • En un mercado con grandes proveedores, como Google o Microsoft, es prácticamente imposible para una pequeña empresa española entrar a competir en el mercado inglés, por lo que se torna esencial abrirse al mercado de la lengua castellana, esto es al mercado latinoamericano. • El castellano es la lengua con mayor índice de aceleración en este tipo de tecnologías: es una lengua que está bien posicionada, hay recursos, hay bases de datos y hay herramientas. Es una lengua conocida que se utiliza que tiene presencia. • La situación de la lengua castellana a nivel internacional mejora respecto a su situación a nivel europeo gracias a la introducción del mercado latinoamericano, que la mayoría de los agentes coinciden en señalar como un nicho de mercado del sector • Se está produciendo un auge de los sistemas neurales para los que se necesita asistencia computacional.

• Existencia de más de 300 investigadores en Iberoamérica, la mayoría en Méjico, con los que poder colaborar para el desarrollo de infraestructuras en la región. • Auge de las redes sociales junto al procesamiento del Big Data que sitúan las industrias de la lengua en una excelente posición para, por un lado, explorar nuevos ámbitos de actuación y por otro, obtener recursos básicos para avanzar en la mejora de los sistemas. • Gran número de agentes implicados en el sector. De manera específica, se identifican necesidades horizontales y globales de esta industria para todas las Administraciones Públicas. • Existencia de programas de doctorado en España con temáticas d procesamiento de lenguaje natural. • Existencia de programas de I+D+i de la UE que pueden ayudar a financiar nuevos proyectos en este campo. • Posibilidad de mejorar la accesibilidad de los colectivos con limitaciones funcionales. • Existencia del Centro de Interoperabilidad Semántica y disponibilidad en el Centro de Transferencia Tecnológica de material para definir soluciones comunes en la Administración.	

En cuarto lugar, se han identificado las amenazas externas del sector de tecnologías del lenguaje español:

- El castellano tiene amplia penetración en distintos países, pero no tiene un predominio tecnológico en lo que respecta a soluciones del lenguaje natural, son los proveedores de tecnologías norteamericanas los que están proporcionando las mejores soluciones de español.
- En la rama de sistemas conversacionales y asistentes de voz, el mercado español en Latinoamérica es una asignatura pendiente, ya que se considera que español es todo y las formas de realización en los distintos países de Latinoamérica son muy diferentes.
- Las empresas españolas compiten con empresas norteamericanas o israelíes con unas legislaciones de protección de datos más flexibles que las europeas y las españolas.

A continuación, se expone la tabla comparativa sobre las amenazas del sector detectadas en el presente estudio y las que se identificaron en 2015:

Amenazas Plan Impulso 2015	Amenazas Estudio TL 2018
• Pérdida de competitividad frente a terceros países, como EE.UU., en el desarrollo de recursos y herramientas para el procesamiento de lenguaje natural y traducción automática para el castellano y español de América. • Falta de acuerdo de estándares y modelos de licencias por parte de diferentes asociaciones y organizaciones europeas que pueden minorar a la industria española si no participa más activamente. • Posibilidad de desaparición del español y lenguas cooficiales como lenguas de dominios especializados si no se fomentan las publicaciones en español y su disponibilidad y uso en el mundo digital. • Requisito de inversión y planificación para ofrecer datos lingüísticos de calidad en portales de datos abiertos, convirtiéndose en una actividad poco sostenible sin financiación específica. • Competencia por parte de grandes empresas en el ámbito de la investigación y desarrollo con los grupos de investigación de procesamiento de lenguaje natural y traducción automática, tanto en español como en lenguas cooficiales.	• El castellano tiene amplia penetración en distintos países, pero no tiene un predominio tecnológico en lo que respecta a soluciones del lenguaje natural, son los proveedores de tecnologías norteamericanas los que están proporcionando las mejores soluciones de español. • En la rama de sistemas conversacionales y asistentes de voz, el mercado español en Latinoamérica es una asignatura pendiente, ya que se considera que español es todo y las formas de realización en los distintos países de Latinoamérica son muy diferentes. • Las empresas españolas compiten con empresas norteamericanas o israelíes con unas legislaciones de protección de datos más flexibles que las europeas y las españolas.

Anexo I. Estado de las Tecnologías del Lenguaje en países próximos de la UE

El presente benchmarking pretende ser un documento de trabajo que sirva para tener una visión general de las estrategias que se están abordando para el impulso del sector, en cuatro países seleccionados: Portugal, Reino Unido, Francia y Alemania.

Partimos del marco europeo como referencia importante a los países de la región, y fuente importante de financiación en tecnologías innovadoras. Posteriormente continuamos realizando un análisis de los países mencionados, donde se recogen las características lingüísticas de cada país, las estrategias o actuaciones lingüísticas y TIC en general, se recogerán las posibles ayudas públicas, en caso de que existan, en cada país que pueda ser del sector de las tecnologías del lenguaje, o en su defecto, que puedan ser adaptable al sector y se acabará recogiendo las materias educativas y las instituciones más relevantes.

Debido al limitado número de informes y estudios relacionado con el sector de las Tecnologías del Lenguaje, este análisis se ha centrado en tres fuentes fundamentales:

- META-NET: Red Europea de Excelencia dedicada a desarrollar los fundamentos tecnológicos de una sociedad europea multilingüe, inclusiva e innovadora.
- LT-Innovate: es la Asociación de la Industria de Tecnología del Lenguaje a nivel internacional.
- Comisión Europea

Si bien la idea original de benchmarking era poder comparar la situación del sector en cuanto a volumen de negocio, principales productos, etc., la bibliografía existente no ofrece dicha información por país, encontrándose cifras desactualizadas o a nivel tecnologías más generales.

Las tecnologías lingüísticas en el contexto europeo

La principal unidad de la Comisión Europea responsable de la tecnología del lenguaje se encuentra en la Dirección G de la Dirección General (DG), denominada "***G3 - Aprendizaje, multilingüismo y accesibilidad***", cuyo objetivo, con respecto al sector de las Tecnologías del Lenguaje, es "*hacer que el mercado único digital sea más accesible, seguro e inclusivo. Con este fin, la unidad respalda la política, la investigación, la innovación y la implementación de las tecnologías de aprendizaje y las tecnologías y servicios clave de lenguaje digital que permitan a todos los consumidores y empresas europeos beneficiarse plenamente del Mercado Único Digital*".

Dentro del marco de las Tecnologías del lenguaje, los principales programas de la Comisión Europea[7] de interés específico para la tecnología del lenguaje son:

- **Horizonte 2020 y su predecesora Programas Marco de Investigación y Desarrollo**. Entre las convocatorias de financiación de la Comisión Europea dentro de este marco se encuentran:
    - H2020 ICT Work Programme 2016 - 2017
        - ICT-14-2016-2017 - Big Data: integración y experimentación de datos multisectoriales e interlingües (8 proyectos)
        - ICT-15-2016-2017 - Big Data: acciones de gran escala piloto en los sectores que mejor se benefician de la innovación basada en datos (2 proyectos)
        - ICT-16-2017 - Big data: investigación que aborda los principales desafíos tecnológicos de la economía de datos
        - ICT-17-2016-2017 - Big data: Apoyo, habilidades industriales, evaluación comparativa y evaluación (1 proyecto)
        - ICT-18-2016 - Big data: tecnologías de big data que preservan la privacidad (4 proyectos)
        - ICT-35-2016 - Permitir la investigación e innovación responsable relacionada con las TIC (1 proyecto: K-PLEX)
    - Programa de trabajo H2020 TIC 2014 - 2015
        - ICT-15-2014 - Big data e Open Data Innovación y aceptación (13 proyectos)
        - TIC-22-2014 - Multimodal e interacción ordenador Natural (2 proyectos Aria-Valuspa y Kristina)
        - ICT-16-2015 - Big data - investigación (10 proyectos)
- El **Programa de Mecanismo de Conexión Europa (CEF) / bloque de construcción en la traducción automática (CEF.AT)**
- **MT@EC**, la traducción automática para las administraciones públicas intiative of DG Translation.

A continuación, pasamos a exponer las principales estrategias lingüísticas de cuatro países de la UE en donde el idioma es referente y es utilizado por una gran parte de población tanto a nivel europeo como

[7] Comisión Europea – Tecnologías del Lenguaje https://ec.europa.eu/digital-single-market/en/language-technologies

mundial. En este sentido, se incluirán los países cuyo idioma está más extendido como son: Reino Unido, Francia, Alemania y Portugal.

En este análisis se realizará un resumen de las estrategias de cada país en relación a las Tecnologías del Lenguaje, si la hubiese, y, del mismo modo, las posibles líneas de financiación que se puedan acoger las empresas, tanto nacionales como europeas, así como los estudios universitarios enfocados al sector y las instituciones públicas nacionales que apoyan el desarrollo de la lengua.

Por último, se recogerá un listado de las principales instituciones que influyan en el sector de las tecnologías del lenguaje en cada uno de los 4 países analizados.

PORTUGAL

Idiomas oficiales

Portugal cuenta con un idioma oficial, el portugués, recogido en el artículo 11 de la "Constitución de la República Portuguesa", y un idioma cooficial, el Mirandese, reconocido en 1999 como idioma cooficial junto al portugués para asuntos locales, y se habla en la ciudad fronteriza nororiental de Miranda do Douro.

A pesar de contar con un idioma cooficial reconocido, Portugal no ha firmado ni ratificado la Carta Europea de Lenguas Regionales o de Minoría[8].

Portugal tiene recogido en su Constitución su idioma oficial. Además, se reconoce un idioma cooficial.

Portugal no ha firmado la Carta Europea de la Lenguas Regionales o de Minoría.

Estrategias y actuaciones lingüísticas

El idioma portugués cuenta con una importante demanda debido, principalmente, por motivos económicos y comerciales relacionados con mercados emergentes como Brasil y Angola.

[8] European Charter for Regional or Minority Languages https://rm.coe.int/168007bf4b

Portugal es miembro y participa en organizaciones creadas para implementar políticas culturales y lingüísticas de la UE, como el diálogo intercultural y el multilingüismo. Concretamente, a través del Instituto Camões, Portugal es uno de los miembros de los Institutos Nacionales de Cultura de la Unión Europea (EUNIC) y la Federación Europea de Instituciones Nacionales de Lenguaje (EFNIL).

Por otro lado, a través de la **Fundación para Ciencia y Tecnología (FCT)** se otorgan fondos a proyectos de investigación específicos del sector, como por ejemplo, se han financiado proyectos relacionados con Corpus de referencia o el proyecto Comprehensive Grammar of Portuguese[9], del Centro de Lingüística de la Universidad de Lisboa.

Otra actuación relevante dentro del sector de las Tecnologías del Lenguaje en Portugal ha sido la inclusión de Portugal como miembro de CLARIN ERIC[10], una red europea de investigación que trabaja en Recursos de lenguaje común e infraestructura de tecnología con el objetivo de crear y mantener una infraestructura para apoyar el intercambio, el uso y la sostenibilidad de los datos de lenguaje y herramientas para la investigación en humanidades y ciencias sociales.

La **Estrategia de investigación e innovación para 2014-2020**[11] es uno de los principales programas de investigación e innovación para el período 2014-2020 en Portugal. Está estrechamente vinculado al Acuerdo de Asociación para los Fondos Regionales y de Inversión Europeos (Fondos EIE) entre Portugal y la UE ("Portugal 2020").

La Estrategia para la Investigación e Innovación para una Especialización Inteligente (IEI) es crucial para la financiación pública de I+D en Portugal ya que presenta como condición previa del Acuerdo de Asociación las prioridades de inversión en investigación e innovación con el ESIF. Las TIC son vitales para fortalecer la cohesión nacional y el desarrollo sostenible.

Esta Estrategia ha sido impulsada por el Ministerio de Economía y el Ministerio de Educación y Ciencia, y en ella se identifican los grandes desafíos en torno a los cuales debe dirigirse la inversión en el período 2014-2020, entre ellos Tecnología para el Idioma Portugués.

[9] Comprehensive Grammar of Portuguese http://www.clul.ulisboa.pt/en/10-research/584-comprehensive-grammar-of-the-portuguese-language

[10] CLARIN ERIC https://www.clarin.eu/news/portugal-joined-clarin-eric

[11] **Estrategia de investigación e innovación para 2014-2020** https://www.portugal2020.pt/Portal2020/programas-operacionais-portugal-2020-2

A pesar de no contar con estrategias enfocadas hacia el sector de las Tecnologías del lenguaje, Portugal tiene varias agencias relevantes y participa activamente en instituciones y asociaciones internacionales relacionadas con las Tecnologías del Lenguaje.

Ayudas públicas nacionales y europeas

Algunos programas potencialmente interesantes para el sector, pero sin un enfoque específico de las Tecnologías del Lenguaje encontramos:

Fondos Nacionales

La **"Fundaçao para a Ciência e a Tecnologia (FCT)"** cuenta con un programa sobre "Desafíos de datos" (*Digging Into Data Challenges*) que está enfocado hacia proyectos de investigación que aborde temas sobre humanidades o ciencias sociales mediante el uso de técnicas de análisis de datos digitales a gran escala y busquen cómo estas técnicas pueden conducir a nuevos conocimientos y enfoques teóricos. Tiene como objetivo avanzar en proyectos colaborativos multidisciplinarios. Los temas más interesantes para los proyectos de Tecnologías del lenguaje se podrían englobar entre las líneas:

1. Interpretar datos;
2. Nueva aplicación de datos;
3. Emplear datos de múltiples fuentes en la investigación.

La financiación nacional para la convocatoria es de 250.000 € y la financiación máxima que se puede solicitar por consorcio con participación portuguesa es de 625.000 €.

Por otro lado, la **Agencia de innovación** cuenta con algunos programas para estimular las inversiones comerciales en el sector de I+D+i, buscando fortalecer las competencias internas de las empresas de I+D en el área de la transferencia de tecnología y el intercambio de conocimientos. Este proyecto tiene como objetivo:

1. El desarrollo de estudios de viabilidad tecnológica;
2. Compartir recursos e infraestructura;
3. Intercambiar recursos humanos calificados entre empresas / organizaciones de I+D con miras a la transferencia de tecnología y el intercambio de conocimientos.

Fondos Europeos

- **Financiación de la ESIF**

Portugal cuenta con diferentes **Programas Regionales**[12]. Aunque no se han encontrado líneas concretas para el sector, si se entiende que existen diversas líneas de financiación que pueden ser potencialmente utilizadas para fortalecer el sector de las tecnologías del lenguaje como el impulso a la investigación y el desarrollo técnico (IDT) y la innovación, así como fomentar la transferencia de conocimientos de IDT e innovación a las Pyme's, apoyar la internacionalización, la competitividad de las empresas y el emprendimiento cualificado.

- Regional OP Alentejo
- Regional OP Algarve
- Regional OP Azores
- Regional OP Centro
- Regional OP Lisboa
- Regional OP Madeira
- Regional OP Norte

Los proyectos del sector de las Tecnologías del Lenguaje se podrían englobar dentro de las Prioridades de la RIS3:

- Las prioridades nacionales de Portugal incluyen Internet y las redes del futuro, la robótica y los sistemas cognitivos, dentro del área de las TIC. Otros sectores que podrían tener relevancia para el sector son el sector turístico, así como el turismo de salud y el sector de la salud.
- **EUREKA / EUROSTARS**

Otras posibles ayudas de financiación pública se encuentran dentro de EUREKA y EUROSTARS, ambas ayudas se ofrecen a través de la Agencia de Innovación y están dirigidas a las universidades e instituciones de investigación, con la diferencia de que, para las ayudas de EUREKA, estas instituciones de investigación deben ir conjuntamente con una empresa.

- **EUREKA:** Universidades e instituciones de investigación solo si van juntas con una empresa.
- **EUROSTARS:** Universidades e instituciones de investigación**.**

[12] Programas regionales
http://ec.europa.eu/regional_policy/en/atlas/programmes?search=1&keywords=&periodId=3&countryCode=PT&regionId=ALL&objectiveId=14&tObjectiveId=ALL

No se han encontrado programas de financiación públicas orientadas principalmente hacia el sector, pero tanto las ayudas, tanto nacionales como internacionales, tienen líneas que pueden ser adaptables para las Tecnologías del Lenguaje.

Educación Universitaria

En Portugal se encuentra el **Programa de Doctorado en Tecnologías del Lenguaje de la Escuela de Ciencias de la Computación de Carnegie Mellon.**

Programa que se centra principalmente en la traducción automática, el procesamiento del habla, y la recuperación de información. Este programa de Doctorado ha sido impulsado por el Instituto de Tecnologías del Lenguaje de Carnegie Mellon con la colaboración del Instituto Técnico Superior de la Universidad Técnica de Lisboa y el Laboratorio NOVA de Ciencias de la Computación e Informática de la Universidad Nova de Lisboa.

Se ha encontrado un programa de Doctorado relacionado con las Tecnologías del Lenguaje, dado las instituciones que colaboran para hacerlo posible, se observa la importancia que está teniendo este sector y el reconocimiento que se le está dando.

Instituciones nacionales que apoyan el desarrollo de la lengua

- **Instituto Internacional de Lengua Portuguesa (Instituto Internacional de Língua Portuguesa - IILP):** Desde 2002, la defensa del idioma portugués y las diferentes culturas de habla portuguesa son sus principales objetivos.
- **Academia de Ciencias de Lisboa:** Contribuir a la promoción del idioma portugués, en particular con la publicación de diccionarios de referencia, el Diccionario de portugués contemporáneo.
- **"Fundaçao para a Ciência ea Tecnologia (FCT)":** es la agencia nacional de financiamiento que apoya la ciencia, la tecnología y la innovación en todos los dominios científicos, bajo la responsabilidad del Ministerio de Educación y Ciencia.

Existen diversas instituciones que apoyan tanto el sector lingüístico como el sector científico que colaboran y promueven proyectos donde el sector de las Tecnologías del lenguaje tiene cabida.

REINO UNIDO

Idiomas oficiales

Dentro del Reino Unido el idioma oficial de facto es el inglés (no existe una constitución formal ni ninguna ley que determine que el inglés es el idioma oficial).

Además, existe una lengua cooficial reconocida, el galés. En 1993 se elaboró la Ley del idioma galés donde se recoge que tanto el inglés como el galés reciban el mismo trato en todo el ámbito del sector público.

Además de estos dos idiomas oficiales, existen otros idiomas minoritarios que se hablan en el Reino Unido y no tienen estatus oficial: el gaélico en Escocia y el irlandés en Irlanda del Norte. Ambos idiomas no tienen el título de idioma cooficial reconocido y en las instituciones públicas y en los centros de enseñanza no se tratan como idioma oficial.

En el caso del gaélico, existe un estatuto especial bajo la ley británica que proporciona ciertas medidas para preservar el idioma. De hecho, la Ley de Lengua Gaélica (Escocia) de 2005 fue aprobada por el Parlamento escocés con el fin de garantizar el estado de la lengua gaélica como idioma oficial de Escocia que impone el mismo respeto al idioma inglés.

Con respecto a los irlandeses, el Reino Unido ha contraído una serie de compromisos vinculantes en relación con el idioma irlandés en Irlanda del Norte bajo la Parte III de la Carta Europea de Lenguas Regionales o Minoritarias.

Por último, hay otro idioma minoritario reconocido por la Carta Europea de Lenguas Regionales o Minoritarias como es el caso del cornish. Este idioma constituye una rama de la sección Celular Insular de la familia de la lengua celta.

Reino Unido cuenta con un idioma oficial, el inglés, el cuál no está recogido en la Constitución, y una lengua cooficial reconocida. Además, existen varias lenguas minoritarias con bastante fuerza dentro del territorio.

Estrategias y actuaciones lingüísticas

Se podría decir que en Reino Unido se está apoyando la investigación de la tecnología del lenguaje desde 1990, cuando se publicó un programa del Departamento de Comercio e Industria sobre

Tecnología del Habla y Lenguaje (SALT), donde se financiaba una amplia gama de pequeños proyectos de colaboración en el campo.

Desde ese período, no se han encontrado programas específicos que implique específicamente el apoyo de la tecnología del lenguaje. Pero si ha habido avances e innovaciones en la introducción del aprendizaje temprano de otros idiomas, en el apoyo a los idiomas de la comunidad y en la promoción de la competencia lingüística para los jóvenes. En parte como resultado de esto, los idiomas permanecen en la agenda política, por ejemplo, en la ***Estrategia de Lenguas Nacionales (2002-2011)*** se sientan las bases para crear un marco para el aprendizaje de idiomas para las edades de siete a once (El marco de Key Stage) y un nuevo marco de evaluación (The Languages Ladder / Asset languages) basado en el MCER (Marco Común Europeo de Referencia para las Lenguas). Las tecnologías del lenguaje no se consideran dentro de estas estrategias pero son parte de los programas de financiación dentro de **EPSRC**[13], la principal agencia del Reino Unido para financiar la investigación en ingeniería y ciencias físicas, incluido el "Procesamiento del lenguaje natural".

Por otro lado, dentro de las lenguas minoritarias, se puede resaltar que:

El gaélico goza de un alto nivel de apoyo político con el plan de idioma gaélico. El primer plan de lenguaje gaélico se definió para el período 2012-2015. Específicamente, dentro de la sección 3, se incluye la subsección "*Corpus de lenguaje*". Para el actual periodo, se lanzó el **Gaelic Language Plan 2015-2020**. En este plan, el Área de Corpus incluye iniciativas centradas en la terminología, la traducción, la ortografía y los topónimos con el fin de garantizar que el gaélico continúe desarrollándose y para lograr una mayor fortaleza, relevancia, consistencia y visibilidad. En ninguno de los planes de lenguaje gaélico, hay una clara referencia al apoyo de las tecnologías del lenguaje.

En relación al galés, existe una estrategia para el periodo anterior, la **Welsh Language Strategy 2012-17,** que tenía entre uno de sus objetivos fortalecer la infraestructura del idioma (incluida la traducción, publicación, investigación, televisión y radio, y las TIC).

Además, Gales ha ofrecido algún tipo de apoyo en el pasado para las tecnologías del lenguaje indirectamente a través de algunos proyectos financiados por FEDER. No se han identificado actividades de traducción automática o reconocimiento de voz significativos dentro de Gales.

[13] EPSRC https://epsrc.ukri.org/

Por último, en Irlanda del Norte dentro del **Programa para el Gobierno 2011-2015**, se incluyó una Estrategia para el idioma irlandés como un componente básico en la Prioridad 4 "*Construir una comunidad sólida y compartida*". A pesar del área de acción en Medios y Tecnología, el apoyo o la promoción de las tecnologías del lenguaje no se menciona en la estrategia.

En ninguno de los planes de lenguaje de Reino Unido y sus regiones, hay una clara referencia al apoyo de las tecnologías del lenguaje, pero se hace referencia al interés de potenciar la lengua en general, la traducción, la terminología y a las TIC e I+D+i.

Ayudas públicas nacionales y europeas

Algunos programas potencialmente interesantes para el sector, pero sin un enfoque de específico de las Tecnologías del Lenguaje encontramos:

Fondos Nacionales

A ámbito nacional se encuentran algunas ayudas enfocadas para Pyme's que podrían considerarse interesantes para el sector. Por ejemplo, en la agencia **MSC R&D**[14] se han encontrado dos programas:

1. **Subvenciones *UK Smart R&D***: Plan enfocado a apoyar a las Pyme's, incluidas las start-ups en las áreas de tecnología y ciencia.
   - Las Pyme's con sede en el Reino Unido pueden solicitar subvenciones de hasta 322,266 euros (£ 250,000) en una base de financiación combinada. Las dos subvenciones principales son:
     - Prueba de concepto: Máximo de 128.915 € (£ 100.000)
     - Desarrollo del prototipo: Máximo de 322.266 € (£ 250.000)
2. **Subvenciones *UK Collaborative R&D***: este plan financia aplicaciones en áreas de tecnologías específicas. Se emite un resumen específico de la competencia y el proceso asociado, los plazos y la financiación para las llamadas temáticas.

   Esta subvención de I+D exige a los solicitantes colabora con otras Pyme's o instituciones académicas y otorga hasta un máximo de 2.552.633 € (£ 2 M).

[14] MSC R&D http://www.mscrnd.com/

Existen otras dos agencias que otorgan subvenciones a nivel nacional, la agencia **Innovate UK Tomorrow**[15] y la agencia **EPSRC/Engineering and Physical Sciences Research Council**[16].

- **IC tomorrow** es un programa de innovación abierto a Pyme's, organizaciones e individuos que buscan oportunidades de financiación para estimular la innovación a través de la tecnología.
- Dentro de la **EPSRC** se apoyan proyectos de la economía digital o las TIC con el objetivo de proporcionar apoyo a corto plazo para permitir que los investigadores de los principales campos de las TIC con agentes de otras disciplinas y/o campos puedan fomentar nuevas colaboraciones e investigaciones, lo que lleva a un enfoque multidisciplinario e impulsado por el usuario a la investigación.

Fondos Europeos

- **Financiación de la ESIF**

Reino Unido tiene 6 **programas operativos regionales**[17]. Entre estos programas no se han encontrado referencias claves para el sector de las TL, pero existen diversas líneas de financiación con los objetivos de fomentar la investigación e innovación. Estos programas permiten a cada Gobierno Regional organizar y prestar una gama de líneas de financiación que apoyan las acciones de innovación y de IDT a lo largo de todo el nivel de la tecnología de la disposición, a partir de la investigación del desarrollo, prueba de concepto hasta la producción, la comercialización y la internacionalización. La ayuda no se limita a las áreas de especialidades clave. Como se puede observar, existe cierta flexibilidad dentro de la gama de programas que se ofrecen para apoyar actividades nuevas y emergentes y que tengan un alto crecimiento o potencial de transformación.

- Reino Unido - FEDER Este de Gales
- Reino Unido - FEDER Inglaterra
- Reino Unido - FEDER Gibraltar
- Reino Unido - FEDER Irlanda del Norte
- Reino Unido - FEDER Escocia

15 Innovate UK Tomorrow https://apply-for-innovation-funding.service.gov.uk/competition/search

16 EPSRC https://epsrc.ukri.org/

17 Programas operativos regionales de Reino Unido http://ec.europa.eu/regional_policy/en/atlas/programmes?search=1&keywords=&periodId=3&countryCode=UK&regionId=ALL&objectiveId=ALL&tObjectiveId=ALL

- Reino Unido - FEDER West Wales and The Valleys
- **EUREKA**

Otras posibles ayudas de financiación pública disponibles son las ayudas EUREKA, esta ayuda está dirigida a las universidades e instituciones de investigación que vayan conjuntamente con una empresa.

No se han encontrado programas de financiación públicas orientadas principalmente hacia el sector, pero tanto las ayudas nacionales como internacionales tienen líneas que pueden ser adaptables para las Tecnologías del Lenguaje.

Educación Universitaria

En Reino Unido existen 3 máster sobre el sector de las tecnologías del lenguaje:

- **Máster en Lingüística Computacional** de la Universidad de Wolverhampton.
- **Máster en Procesamiento de Habla y Lenguaje** en la Facultad de Filosofía, Psicología y Ciencias del Lenguaje de la Universidad de Edimburgo.
- **Máster en Ciencias de la Computación con Procesamiento de Habla y Lenguaje** en la Facultad de Ciencias de la Computación de la Universidad de Sheffield.

Se ha encontrado 3 máster relacionado con las Tecnologías del Lenguaje, y posiblemente no sean los únicos que existan, lo que resalta la importancia que está teniendo este sector y el reconocimiento que se le está dando a nivel nacional.

Instituciones nacionales que apoyan el desarrollo de la lengua

Gaélico:

- **Gobierno escocés**: aprendizaje, ciencia e idiomas de Escocia.
- **Bòrd na Gàidhlig:** Es el organismo público ejecutivo no departamental del Gobierno escocés con responsabilidad gaélica.

Galés:

- **The Partnership Council**: es responsable de dar consejos y hacer representaciones a los Ministros en relación con la estrategia del idioma galés.
- **Gobierno de Gales:** Unidad de idioma galés.
- **Coleg Cymraeg Cenedlaethol:** National Welsh Language College se estableció en 2011.

Irlandés:

- **Foras na Gaeilge**: El organismo responsable de la promoción del idioma irlandés en toda la isla de Irlanda.

Existen diversas instituciones a nivel regional que apoyan tanto el sector lingüístico como el sector científico que colaboran y promueven proyectos donde el sector de las Tecnologías del lenguaje tiene cabida.

FRANCIA

Idiomas oficiales

La Constitución francesa, en su Título 1, Artículo 2 establece que "*el idioma de la República será el francés*". Este artículo, y la centralización de los diferentes gobiernos, han hecho que se repriman los posibles idiomas regionales o minoritarios.

Actualmente, no hay idiomas reconocidos como idiomas cooficiales. Esto no significa que no hay ninguno. Un informe de 1999 identificó 75 idiomas en Francia que calificarían como idioma regional o minoritario bajo la Carta Europea de Lenguas Regionales o Minoritarias.

- Bretón: lengua céltica. Noroeste de Francia.
- Corso: dialecto toscano. Córcega.
- Provenzal: dialecto del occitano. Provenza, Languedoc.
- Occitano: Lengua romance. Al sur del Loire (Niza, Valle de Arán, Bearne, Aude)
- Catalán: proviene del latín gálico. Pirineos Orientales.
- Franco-provenzal: lengua románica. Savoie, Fribourg, y Valais.
- Alsaciano: dialecto de origen germánico. Alsacia.
- Vasco: Pirineos Atlánticos (País Vasco-Francés)

Durante el mandato de François Hollande se buscó generar un marco legal claro para los idiomas regionales dentro de un programa de descentralización administrativa que otorgaría competencias a

las regiones en materia de política lingüística. Pero a finales de julio de 2015, el Consejo Constitucional frenó dicho cambio constitucional.

Francia cuenta con un idioma oficial, el francés, el cual está recogido en la Constitución, pero no tiene ninguna lengua cooficial reconocida.

Estrategias y actuaciones lingüísticas

Francia introdujo en 1994 una ley para la preservación del idioma francés contra los anglicismos, lo que hace que, en los textos públicos, la televisión o la publicidad no se puedan usar palabras de origen inglés. La "***Delegación general para la lengua francesa y los lenguajes de Francia (DGLFLF)***" sustituye las palabras en inglés por palabras francesas.

En 2002 se publicó el **programa Techno-langue**, un gran programa nacional francés sobre tecnologías lingüísticas que duró hasta 2006. En la página web del proyecto aún se puede encontrar la lista de los proyectos que fueron financiados. Por ejemplo, se financió el proyecto CESART que permitió llevar a cabo una campaña para la evaluación de herramientas de extracción terminológica y traducción automática.

En 2006, el gobierno francés encargó un estudio sobre "Tecnologías del lenguaje en Europa". Una de sus conclusiones fueron que "*Varios factores incitan a los responsables de la toma de decisiones a integrar soluciones innovadoras en su empresa para gestionar de forma inteligente el contenido digital [...] el uso progresivo de las TIC nos permite predecir que el mercado de herramientas lingüísticas se abrirá hacia el público en general. Se siente la necesidad de tomar medidas de marketing para optimizar la oferta y la demanda*."

Recientemente, Francia ha publicado la "***Guide des bonnes pratiques linguistiques dans les emprende",*** emitida por DGLFLF, esta guía está dirigida a las empresas francesas que trabajan a nivel internacional para conciliar el uso del francés con la necesidad de una comunicación global. Del mismo modo, La DGFLFL emitió en 2014 un documento resumen sobre "***Las tecnologías digitales al servicio de los idiomas***" en el que señalan iniciativas que se pretenden llevar a cabo a corto y largo plazo:

A corto plazo:

- Crowdsourcing para enriquecer el idioma francés.
- "*JocondeLab*": Proyecto que pretende traducir en 14 idiomas casi 300.000 obras de arte.

- *Iniciativa SémanticPédia*: una colaboración del Ministerio de Cultura y Comunicación, el *Institut National de Recherche en Informatique et en Automatique* (INRIA) y la Wikimédia France. Un taller de un día (cooperación del CNRS y el Ministerio de Investigación) como un trampolín para un gran programa nacional para apoyar el desarrollo de herramientas lingüísticas para el francés y las lenguas minoritarias y regionales de Francia, citadas al principio.

A largo plazo:

- Se prevé una mayor ampliación de la experiencia de SémanticPédia a largo plazo. La creación de un "Technolangue II" que podría ser financiado por el “Program d'Investissements d'Avenir (PIA)”.
- Además, se prevé una cooperación más estrecha con la UE, en particular en las áreas de cultura (EUROPEANA), aprendizaje (DG EAC) y tecnologías del lenguaje.

Actualmente, el presidente Emmanuel Macron está decidido a que Francia vuelva a ser un actor decisivo en el ajedrez internacional. Y lo quiere hacer en todos los ámbitos, no solo en la política. En este sentido, el Macron anunció que, hasta 2022, se destinarán 1.500 millones de euros de fondos públicos al Plan de Inteligencia Artificial para promover investigación y proyectos en la materia. Además, aseguró que la IA será “el primer campo de aplicación” del Fondo para la Innovación y la Industria, con 10.000 millones de euros, lanzado a comienzos de año.

En ninguno de los planes de lenguaje de Francia hay una clara referencia al apoyo de las tecnologías del lenguaje, se observa que se está invirtiendo bastante dinero en temas TIC e inteligencia artificial, líneas adaptables para el sector.

Ayudas públicas nacionales y europeas

Algunos programas potencialmente interesantes para el sector, pero sin un enfoque específico de las Tecnologías del Lenguaje encontramos:

Fondos Nacionales

A nivel nacional existen diversos programas de financiamiento, pero por lo general suelen estar disponibles en forma de préstamos. También existen subvenciones reembolsables, pero solo cuando tienen éxito.

Los programas de la Agencia **ANR/ The French National Research Agency**[18] se enmarcan dentro del ámbito de la Agenda Estratégica de Investigación y Transferencia e Innovación "Francia Europa 2020".

- *Digging into* Data: el objetivo de este programa transatlántico es apoyar proyectos que usan "big data" para abordar preguntas en ciencias sociales y humanidades. Los proyectos pueden proponer desarrollar nuevas herramientas, aplicaciones y métodos mediante el uso de técnicas de análisis de datos digitales y mostrar cómo estas técnicas pueden conducir a nuevos conocimientos.
- *"Setting Up European or International Scientific Networks (MRSEI)":* el objetivo de este programa es facilitar el acceso de los investigadores franceses a los programas europeos y otras convocatorias internacionales, así como fomentar la formación y coordinación de redes transnacionales. Se proporcionará financiación en todas las disciplinas para redes de investigación específicamente destinadas a la preparación y presentación de proyectos de colaboración, tanto europea como internacional, con gran impacto tecnológico y científico.

Fondos Europeos

- **Financiación de la ESIF**

En cuanto a los **programas operativos y regionales existentes**[19] en Francia, encontramos que cuenta con 2 programas operativos, 22 programas regionales y un programa nacional de asistencia. En estos programas, al igual que ocurre en los demás países, existen líneas de financiación para fomentar entre otras actuaciones, la investigación e innovación tecnológica.

- Programa operativo ERDF-ESF Guadalupe y San Martín Etat 2014-2020
- Programa Operativo ERDF-ESF ile-de-France et Seine 2014-2020
- Programa nacional de asistencia técnica 2014-2020
- Programa regional Aquitaine 2014-2020
- Programa regional Auvernia 2014-2020
- Programa regional Basse-Normandie 2014-2020

[18] ANR http://www.agence-nationale-recherche.fr/

[19] Programas Operativos y Regionales en Francia http://ec.europa.eu/regional_policy/en/atlas/programmes?search=1&keywords=&periodId=3&countryCode=FR&regionId=ALL&objectiveId=ALL&tObjectiveId=ALL#

- Programa regional Bretagne 2014-2020
- Centro de programa regional 2014-2020
- Programa regional Champagne-Ardenne 2014-2020
- Programa regional Corse 2014-2020
- Programa regional Franche-Comté et Jura 2014-2020
- Programa regional Guadalupe Conseil Régional 2014-2020
- Programa regional Guyane Conseil Régional 2014-2020
- Programa regional Alta Normandía 2014-2020
- Programa regional Languedoc-Roussillon 2014-2020
- Programa regional Limousin 2014-2020
- Programa regional Martinica Conseil Régional 2014-2020
- Programa regional Mayotte 2014-2020
- Programa regional Midi-Pyrénées et Garonne 2014-2020
- Programa regional Nord-Pas de Calais 2014-2020
- Programa regional Pays de la Loire 2014-2020
- Programa regional Picardie 2014-2020
- Programa regional Poitou Charentes 2014-2020
- Programa regional Provence Alpes Côte d'Azur 2014-2020
- Programa regional Rhône Alpes 2014-2020

Las oportunidades de ESIF están disponibles a nivel regional, por ejemplo, Nord-Pas-de-Calais, Lorraine y Bourgogne, donde el sector de las Tecnologías del Lenguaje tendría una oportunidad dentro de las Prioridades de la RIS3.

- **EUREKA / EUROSTARS**

Otras posibles ayudas de financiación pública se encuentran dentro de EUREKA y EUROSTARS, ambas ayudas están dirigidas a las universidades e instituciones de investigación, con la diferencia que para las ayudas de EUREKA, estas instituciones de investigación deben ir conjuntamente con una empresa.

- **EUREKA:** Universidades e instituciones de investigación solo si van juntas con una empresa.
- **EUROSTARS:** Universidades e instituciones de investigación**.**

No se han encontrado programas de financiación públicas orientadas principalmente hacia el sector, pero tanto las ayudas nacionales como internacionales tienen líneas que pueden ser adaptables para las Tecnologías del Lenguaje.

Educación Universitaria

No se ha encontrado máster o programas educativos relacionados con las Tecnologías del Lenguaje en Francia.

Instituciones nacionales que apoyan el desarrollo de la lengua

- **Ministry of Culture - General Delegation for the French Language and the Languages of France**: Su misión es liderar, a nivel interdepartamental, la política lingüística de Francia.

Se ha encontrado una institución del Gobierno desde donde se lideran las políticas lingüísticas a nivel internacional.

ALEMANIA

Idiomas oficiales

El idioma oficial de Alemania es el alemán. Sin embargo, esto no está establecido en la constitución, sino solo en las leyes federales / regionales. Han existido diversas iniciativas para cambiar la constitución e incluir el alemán como idioma oficial, pero hasta la fecha todas estas iniciativas no han salido adelante.

Por otro lado, Alemania no tiene idiomas cooficiales, pero firmó y ratificó la Carta Europea de Lenguas Regionales o Minoritarias debido a que sí reconoce algunos idiomas regionales: Sorbian, Romani, Danés y North Frisian.

Alemania cuenta con un idioma oficial, el alemán, el cuál no está recogido en la Constitución.
En Alemania no existe una lengua cooficial reconocida, pero si se reconocen diversos idiomas minoritarios.

Estrategias y actuaciones lingüísticas

En Alemania, la competencia lingüística la tienen los Laender (estados constituyentes) y, a excepción de la reforma del idioma alemán de 1996 ("Rechtschreibreform"), no se ha llevado a cabo ninguna iniciativa con respecto a los idiomas a nivel federal.

A nivel regional, las políticas lingüísticas se refieren principalmente a la educación escolar. De 38.000 escuelas en Alemania, solo 200 son bilingües y estas son principalmente alemán-inglés o alemán-francés.

No se ha encontrado ninguna estrategia de lenguaje público o enfocado a las Tecnologías del Lenguaje.

Ayudas públicas nacionales y europeas

Algunos programas potencialmente interesantes para el sector, pero sin un enfoque específico de las Tecnologías del Lenguaje encontramos:

Fondos Nacionales

La agencia **Förderberatung des Bundes**[20] ofrece asesoramiento sobre programas de I+D+i del gobierno alemán. Según la información recabada, el panorama de I+D+i en Alemania es complejo debido a la competencia a nivel federal y estatal, y suele ser complicado acceder a ayudas, por eso, esta agencia brinda información y consejos para encontrar las ayudas deseadas, así como de los caminos a seguir para solicitarlas.

Entre los principales temas de financiación[21] se incluyen:

- **TIC 2020 - Investigación para la innovación:** Las tecnologías de la información y la comunicación como impulsoras de la innovación.
- **Tecnologías digitales:** Tecnologías de información y comunicación con alto potencial de aplicación y transferencia.
- **De seguridad de TI: La seguridad de TI garantiza la confidencialidad, la integridad y la disponibilidad de la información y la tecnología de la información.**
- **Interacción hombre-máquina:** Adaptar de forma óptima las tecnologías modernas a las necesidades de **las** personas.
- **Microelectrónica:** Microelectrónica como impulsor de la innovación para la economía y la sociedad.

[20] Förderberatung des Bundes https://www.foerderinfo.bund.de/

[21] https://www.foerderinfo.bund.de/de/kommunikation-191.php

- **Medios digitales:** Herramientas de aprendizaje y trabajo basadas en computadora en la educación vocacional.

Fondos Europeos

- **Financiación de la ESIF**

En cuanto a los **programas operativos y regionales existentes**[22] en Alemania, encontramos que cuenta con 16 programas operativos para el periodo actual. En estos programas, y similar a los demás países, las principales líneas de financiación que podrían estar relacionadas con el sector de las tecnologías del lenguaje se encuentran entre los objetivos de fomentar la investigación e innovación tecnológica.

- OP Baden-Württemberg FEDER 2014-2020
- OP Bayern FEDER 2014-2020
- OP Berlín FEDER 2014-2020
- OP Brandeburgo FEDER 2014-2020
- OP Bremen FEDER 2014-2020
- OP Hamburgo FEDER 2014-2020
- OP Hessen FEDER 2014-2020
- OP Mecklemburgo-Pomerania Occidental FEDER 2014-2020
- OP Niedersachsen FEDER / FSE 2014-2020
- OP Nordrhein-Westfalen FEDER 2014-2020
- OP Rheinland-Pfalz FEDER 2014-2020
- OP Saarland FEDER 2014-2020
- OP Sachsen FEDER 2014-2020
- OP Sachsen-Anhalt FEDER 2014-2020
- OP Schleswig-Holstein FEDER 2014-2020
- OP Thüringen FEDER 2014-2020

Las posibles ayudas de ESIF con respecto al sector estarían disponibles a nivel regional, por ejemplo, Bayern o Berlín, donde las Tecnologías del lenguaje podrían adaptarse dentro de las Prioridades de la RIS3.

[22] Programas Operativos y Regionales en Alemania http://ec.europa.eu/regional_policy/en/atlas/programmes?search=1&keywords=&periodId=3&countryCode=DE&regionId=ALL&objectiveId=ALL&tObjectiveId=ALL

- **Prioridades del de la RIS3 para el Bayern:** Las tecnologías del lenguaje están cubiertas por la estrategia de innovación regional de Baviera para la especialización inteligente.
- **Prioridades del de la RIS3 para el Berlín:** De acuerdo con información, el estado federal de Berlín concede subvenciones o préstamos, principalmente para pequeñas y medianas empresas, pero también para institutos de investigación, para proyectos en las áreas de investigación industrial, desarrollo experimental y desarrollo de producción, preparación de mercado y lanzamiento. El enfoque del financiamiento público se basa en los siguientes campos tecnológicos:
    - Información y comunicación, medios e industrias creativas,
    - Economía de la salud (atención médica, telemedicina, turismo de salud, etc.),
    - Transporte, movilidad y logística (telemática del tráfico, automoción, tecnología de transporte ferroviario, aeroespacial, etc.).
- **EUREKA / EUROSTARS**

Otras posibles ayudas de financiación pública se encuentran dentro de EUREKA y EUROSTARS, ambas ayudas están dirigidas a las universidades e instituciones de investigación, con la diferencia que para las ayudas de EUREKA, estas instituciones de investigación deben ir conjuntamente con una empresa.

- **EUREKA:** Universidades e instituciones de investigación solo si van juntas con una empresa. Dentro de EUREKA se encuentran dos programas específicos:
    - ZIM: programa de innovación para Pyme's e institutos de investigación/universidades ofrecido a través de la Agencia de gestión de proyectos de BMWi.
    - Centro Aeroespacial Alemán (DLR): todos los beneficiarios.
- **EUROSTARS:** Universidades e instituciones de investigación**.**

Dentro de la agencia Förderberatung des Bundes se encuentra una línea relacionada con el sector, lo que da a entender que posiblemente existan líneas o programas que sean específicos o estén enfocado al sector.

Educación Universitaria

Universidades con estudios especializados en tecnologías del lenguaje:

- *"Master Language Science and Technology"* impartido en la Universidad de Saarbrücken junto al Centro Alemán de Investigación sobre Inteligencia Artificial (DFKI);

- *"Master`s course in Language Technology and Foreign Language Teaching"* de la Justus-Liebig-Universität Gießen,
- La Heinrich-Heine Universität Düsseldorf cuenta con un Departamento de Informática Lingüística y tienen títulos enfocados a las Ciencia de la información y tecnología del lenguaje,
- El Technische Hochschule Colonia cuenta con un programa de Máster en Terminología y Tecnología del Lenguaje.

En las universidades alemanas existen diversos estudios de máster y títulos propios relacionados con las tecnologías del lenguaje. Se puede observar la importancia este sector en el ámbito académico alemán ya que en muchas de sus universidades se impartes estudios específicos para el sector, proporcionado así una mayor especificidad y profesionalidad en el ámbito laboral.

Instituciones nacionales que apoyan el desarrollo de la lengua

- La **Sociedad Alemana de Lingüística Computacional y Tecnología del Lenguaje (GSCL):** es la asociación científica para la investigación, la enseñanza y el trabajo profesional en procesamiento del lenguaje natural. Es compatible con la cooperación con disciplinas vecinas (por ejemplo, lingüística y semiótica, computadora ciencia de la información y la ciencia, ciencia psicológica y ciencia cognitiva) y mantiene contacto con las asociaciones respectivas.
- **Centro Alemán de Competencia en Tecnología del Habla y el Lenguaje** en el Centro Alemán de Investigación de Inteligencia Artificial (DFKI).
- **Goethe-Institut**: Instituto mundial de promoción y enseñanza del idioma alemán.
- **Gesellschaft für deutsche Sprache (DfdS):** La sociedad del idioma alemán no tiene fines de lucro asociación para la protección e investigación del idioma alemán.
- **Institut für deutsche Sprache (Instituto para el Idioma Alemán):** Este instituto de Mannheim investiga el idioma alemán y su uso e historia reciente.
- **Verein Deutsche Sprache (VDS):** Asociación sin fines de lucro que protege y promueve el idioma alemán. Controlador principal para incluir el idioma alemán en la constitución.

Existen diversas instituciones a nivel nacional que apoyan tanto el sector lingüístico como el sector científico que colaboran y promueven proyectos donde el sector de las Tecnologías del lenguaje tiene cabida.

RESUMEN

Como se ha podido observar, las tecnologías del lenguaje no desempeñan actualmente un papel ni en la agenda política europea, ni en los países analizados, y tampoco se reflejan adecuadamente en las políticas actuales de los países ni de la UE sobre tecnologías de la información y la comunicación.

Aunque existe una gran diversidad de idiomas dentro de cada país, las herramientas y los recursos para potenciar el sector y buscar nuevas formas de potenciar la comunicación son escasos, en algunos casos casi inexistentes. La mayor parte de las ayudas y programas públicos están enfocados a las TIC, la Inteligencia Artificial, el big Data, etc., dejando a las empresas y centros de investigación del sector a merced de competir dentro de estas ayudas sin ninguna especifidad.

En algunos casos, como en Alemania, por ejemplo, se observa que se está apostando por la gestión documental y los interfaces humano-maquina. Portugal se está enfocando al Procesamiento del Lenguaje Natural, y a pesar de no encontrar datos concretos, el mercado de la traducción automática se entiende que está creciendo a consecuencia de las actuaciones enfocadas fomentar el lenguaje y la traducción.

El caso de los recursos de lenguaje, los corpus y la terminología, son materias de Tecnología del Lenguaje que se están desarrollando en los cuatro países, y en general, en Europa.

Anexo II. Guía de oportunidad de financiación e inversión

La financiación de I+D puede proceder, en general, de cinco tipos de fondos: las administraciones públicas (estatal, autonómica y local), la enseñanza superior (universidades), las empresas, las instituciones privadas sin finalidad de lucro (IPSFL) y el extranjero (aquellas aportaciones que provienen de fuera del Estado y que normalmente suelen estar lideradas por los programas europeos).

Para elaborar esta guía, las líneas de financiación e inversión detectadas se han agrupado en tres categorías:

- Fondos públicos: incluyen administraciones públicas y enseñanza superior.
- Fondos privados: incluyen empresas e IPSFL
- Fondos provenientes del extranjero

Fondos públicos

Organismo	Programa/línea de financiación	Descripción

Centro para el desarrollo tecnológico industrial (CDTI)	Proyectos de investigación y desarrollo	Están dirigidos a empresas y su objetivo es la financiación de proyectos de I+D desarrollados por empresas y destinados a la creación y mejora significativa de procesos productivos, productos o servicios.
		Los proyectos de I+D son proyectos orientados a la creación y/o mejora significativa de un proceso productivo, producto o servicio que pueden comprender tanto actividades de investigación industrial como de desarrollo experimental. No existe ninguna restricción en cuanto al sector o a la tecnología a desarrollar.
		Se distinguen tres categorías de proyectos:
		• Proyectos de I+D individuales.
		• Proyectos de I+D en cooperación nacional.
		• Proyectos de cooperación tecnológica internacional: estos proyectos tienen como objetivo fomentar la cooperación tecnológica con entidades de otros países o bien capacitar a empresas españolas para mejorar su participación en programas internacionales, concretamente:
		  ◦ Proyectos promovidos por consorcios internacionales o relacionados con la participación española en programas de cooperación tecnológica internacional gestionados por el CDTI (programas multilaterales y bilaterales, programa de proyectos internacionales con certificación y seguimiento unilateral por CDTI, y proyectos ERANETs).
		  ◦ Proyectos relacionados con el incremento de la capacidad tecnológica de las empresas españolas para mejorar su posible participación en Proyectos Importantes de Interés Común Europeo (PIICE). Los PIICE son grandes iniciativas que contribuyen a la realización de los objetivos de la Unión Europea, con los que participan varios Estados

Organismo	Programa/línea de financiación	Descripción
		miembros y de importancia cuantitativa o cualitativa, tal y como se definen en la Comunicación de la Comisión "Criterios para el análisis de la compatibilidad con el mercado interior de las ayudas para fomentar la realización de proyectos importantes de interés común europeo" (2014/C 188/02). • Proyectos de Iniciativas Tecnológicas Conjuntas (JTI, en inglés) en las que no se financie la participación de la gran empresa, concretamente la JTI IMI (Innovative Medicines Initiative) y la JTI BBI (Bio-based industries). Las Iniciativas Tecnológicas Conjuntas son asociaciones público-privadas que se centran en ámbitos esenciales donde la investigación y la innovación puedan contribuir a los objetivos de competitividad generales de la Unión y a la solución de sus retos sociales y en las que es necesario incentivar la participación de grandes empresas españolas como tractoras en estas iniciativas, dado que están excluidas de financiación europea. • Proyectos relacionados con el incremento de la capacidad tecnológica de las empresas españolas para mejorar su posible participación en las Grandes Instalaciones Científico-Tecnológicas Internacionales.
Centro para el desarrollo tecnológico industrial (CDTI)	Proyectos estratégicos CIEN	El Programa Estratégico de Consorcios de Investigación Empresarial Nacional (CIEN) financia grandes proyectos de investigación industrial y de desarrollo experimental, desarrollados en colaboración efectiva por agrupaciones empresariales y orientados a la realización de una investigación planificada en áreas estratégicas de futuro y con potencial proyección internacional. Persigue, además, fomentar la cooperación público-privada en el ámbito de la I+D por lo que requiere la subcontratación relevante de actividades a organismos de investigación. Las actividades de investigación industrial y de desarrollo experimental son las definidas en la normativa europea sobre ayudas de estado.
Centro para el desarrollo tecnológico industrial (CDTI)	Proyectos de innovación CDTI	Están dirigidos a empresas y su objetivo es la financiación de proyectos que permitan la incorporación y adaptación de tecnologías novedosas a nivel sectorial, y cuya implantación represente una ventaja competitiva para la empresa. Apoyo a empresas con proyectos de innovación tecnológica con alguno de los siguientes objetivos: • Incorporación y adaptación activa de tecnologías que supongan una innovación en la empresa, así como los procesos de adaptación y mejora de tecnologías a nuevos mercados. • Aplicación del diseño industrial e ingeniería de producto y proceso para la mejora tecnológica. • Aplicación de un método de producción o suministro nuevo o significativamente mejorado. No existe ninguna restricción en cuanto a sector o tecnología.
Centro para el desarrollo tecnológico industrial (CDTI)	Proyectos de innovación global	Dirigidos a empresas pequeñas y medianas (pymes) y de mediana capitalización (midcaps), cuyo objetivo es la financiación de proyectos de inversión en innovación que permitan la internacionalización y crecimiento empresarial de los beneficiarios.

Organismo	Programa/línea de financiación	Descripción
		Estas ayudas están destinadas a financiar proyectos de inversión en innovación y destinados a la incorporación de tecnología innovadora para la internacionalización y crecimiento empresarial de empresas que desarrollen sus actividades en España, incluyendo las instalaciones ubicadas en el extranjero. Los proyectos han de estar dirigidos a la incorporación de tecnologías necesarias para la adaptación a nuevos mercados, mejorar la posición competitiva de la empresa y contribuir a la generación de valor añadido.
Centro para el desarrollo tecnológico industrial (CDTI)	Proyectos innoglobal	El programa Innoglobal está diseñado para impulsar la cooperación tecnológica internacional de las empresas españolas. Se subvencionará a las empresas españolas presentes en proyectos dentro de los Programas Multilaterales (EUREKA e IBEROEKA), Bilaterales (Japón, China, India, Brasil, Rusia…) o en los proyectos internacionales de certificación unilateral por CDTI, junto a otros relativos a su preparación para la participación en licitaciones de organismos de Investigación y Grandes Instalaciones Científicas internacionales, de acuerdo con las prioridades establecidas en el Programa Estatal de Impulso al Liderazgo Empresarial en I+D. El objetivo principal de esta ayuda es apoyar la acción internacional de las empresas españolas ejerciendo un efecto catalizador capaz de incrementar el número de proyectos de cooperación internacional cercanos a mercado, que movilicen la inversión privada, generan empleo y ayuden a mejorar la balanza tecnológica del país.
Centro para el desarrollo tecnológico industrial (CDTI)	Proyectos Neotec	El programa NEOTEC tiene como objetivo el apoyo a la creación y consolidación de empresas de base tecnológicas. Las ayudas del programa financian la puesta en marcha de nuevos proyectos empresariales que requieran el uso de tecnologías o conocimientos desarrollados a partir de la actividad investigadora, en los que la estrategia de negocio se base en el desarrollo de tecnología. La tecnología y la innovación han de ser factores competitivos que contribuyan a la diferenciación de la empresa y que sirvan de base a la estrategia y al plan de negocio a largo plazo, con el mantenimiento de líneas de I+D propias. Las ayudas podrán destinarse a proyectos empresariales de cualquier ámbito tecnológico y/o sectorial. El beneficiario debe ser una pequeña empresa innovadora.
Centro para el desarrollo tecnológico industrial (CDTI)	Programa Innvierte	El programa INNVIERTE forma parte de la Estrategia Española de Ciencia y Tecnología y de Innovación 2013-2020. Esta estrategia contiene los objetivos, las reformas y las medidas que deben adoptarse en todo el ámbito de la I+D+i con el fin de impulsar su crecimiento e impacto, y es uno de los pilares sobre los que se asiente el diseño de la política del Gobierno en I+D+i para los próximos años. El programa INNVIERTE persigue promover la innovación empresarial mediante el apoyo a la inversión de capital riesgo en empresas de base tecnológica o innovadoras.
Ministerio de Ciencia, Innovación y Universidades	Horizonte pyme	Se plantea como una segunda oportunidad para aquellas startups que se han quedado fuera de los programas Horizonte 2020. La Comisión Europea avala y da el visto bueno a muchas startups pero no puede financiarlas a todas, ahí es donde entra el programa

Organismo	Programa/línea de financiación	Descripción
		Pyme, que financia estudios de viabilidad (técnicos y comerciales) que incluyan un plan de negocio de un proyecto innovador a aquellas pymes que han sido evaluadas por la CE con una puntuación igual o superior a 12 puntos.
Ministerio de Ciencia, Innovación y Universidades	Emplea	Esta convocatoria de ayudas tiene como objetivo incentivar el desarrollo de actividades I+D+I en las pequeñas y medianas empresas y la creación de empleo de calidad para titulados universitarios y titulados en formación profesional de grado superior o equivalente que realicen actividades de I+D+I en pymes, spin-off o Joven Empresa Innovadora (JEI).
ENISA	Línea ENISA emprendedores	El objetivo de esta línea es apoyar financieramente en las primeras fases de vida a pymes promovidas por emprendedores para que acometan las inversiones necesarias y lleven a cabo su proyecto. Las ayudas están dirigidas a pymes con un modelo de negocio innovador/novedoso o con claras ventajas competitivas.
ENISA	Línea ENISA crecimiento	El objetivo de esta línea es financiar proyectos basados en modelos de negocio viables y rentables, enfocados a una mejora competitiva de sistemas productivos y/o cambio de modelo productivo; expansión mediante ampliación de la capacidad productiva, avances tecnológicos, aumento de gama de productos/servicios, diversificación de mercados…; búsqueda de capitalización y/o deuda en mercados regulados y financiación de proyectos empresariales a través de operaciones societarias. Las ayudas están dirigidas a pymes que contemplen mejoras competitivas, proyectos de consolidación, crecimiento e internacionalización y operaciones societarias.
Agencia IDEA (Agencia de Innovación y Desarrollo de Andalucía)	Programa de incentivos para la promoción de la investigación industrial, el desarrollo experimental y la innovación empresarial en Andalucía	El objetivo de esta orden es el incremento de la competitividad de las empresas a través de la generación e incorporación de tecnologías e innovaciones destinadas a la mejora de procesos y la creación de productos y servicios tecnológicamente avanzados y de mayor valor añadido. En el marco de esta orden, se financian actuaciones que se enmarquen dentro de alguno de los siguientes programas: • Programa de apoyo a la I+D+i empresarial • Programa de fomento de la I+D+i internacional • Programa de liderazgo en innovación abierta, estratégica y singular
Agencia IDEA (Agencia de Innovación y Desarrollo de Andalucía)	Programa de incentivos para el desarrollo industrial, la mejora de la competitividad, la transformación digital y la creación de empleo	El objetivo de esta orden es contribuir al desarrollo industrial y a la creación de empleo, mediante la mejora de la competitividad de las empresas o fomentando la creación o el crecimiento de empresas generadoras de empleo. De igual forma, tiene como objetivo el impulso de la innovación productiva en los ámbitos de la especialización inteligente y la incorporación de servicios avanzados para la gestión empresarial, la dinamización empresarial y la cooperación. En el marco de esta orden, se financian actuaciones dentro de alguna de las siguientes líneas: • Creación de actividad económica • Mejora de la competitividad empresarial • Generación de empleo • Servicios avanzados

Organismo	Programa/línea de financiación	Descripción
		• Transformación digital de las pymes
Instituto Vasco de Finanzas	Programa de apoyo financiero a la I+D+i, a la inversión en medidas de eficiencia energética y energías limpias y a la inversión científico-tecnológica	Su objetivo es fomentar la realización de proyectos en investigación, desarrollo e innovación, en medidas de eficiencia energética y energías limpias, y en inversiones científico-tecnológicas de empresas del sector industrial y servicios conexos, y Centros Tecnológicos, Centros de Investigación Cooperativa (Cic,s) y Unidades de I+D empresariales adscritos a la RVCTI, mediante préstamos a conceder por parte del Instituto Vasco de Finanzas en colaboración con las Entidades Financieras, y a través de la prestación de garantías a dichas operaciones de préstamo.
Xunta de Galicia Consellería de Economía Emprego e Industria	Fondo Galicia Iniciativas Emprendedoras (FGIE)	Fondo de inversión constituido por el Igape, Xesgalicia y Gain, cuyo objetivo es contribuir al fomento del espíritu emprendedor y dar apoyo financiero a las iniciativas emprendedoras.

Fondos provenientes del extranjero

Organismo	Programa/línea de financiación	Descripción
Centro para el desarrollo tecnológico industrial (CDTI)	Proyectos FEDER innterconecta	El CDTI como como gestor de FEDER a partir de la ronda 2007-2013, diseñó un instrumento de carácter regional para potenciar la generación de capacidades innovadoras en las regiones menos desarrolladas a través de la financiación de proyectos de desarrollo experimental realizados mediante consorcios empresariales: FEDER Innterconecta, dirigidos a Andalucía, Canarias, Castilla La Mancha, Extremadura, Galicia, Murcia, Ceuta y Melilla. Las temáticas de los proyectos deberán responder a alguno de los ocho Retos Sociales establecidos en el artículo 8.2 c) de la Orden de bases reguladoras: • Salud, cambio demográfico y bienestar. • Seguridad y calidad alimentarias; actividad agraria productiva y sostenible recursos naturales, investigación marina y marítima. • Energía segura, eficiente y limpia. • Transporte inteligente, sostenible e integrado. • Acción sobre el cambio climático y eficiencia en la utilización de recursos y materias primas. • Cambios e innovaciones sociales. • Economía y sociedad digital. • Seguridad, protección y defensa. Mediante este instrumento, el CDTI pretende impulsar la cooperación en el ámbito regional, la realización de proyectos orientados a las necesidades de las regiones y la generación de capacidades innovadoras que fomenten una mayor cohesión territorial. Las convocatorias de FEDER Innterconecta cuentan con la cofinanciación de FEDER a través de los distintos Programas Operativos en los que el CDTI ha sido Organismo Intermedio. En la ronda 2014-2020, las convocatorias se cofinanciarán a través del Programa Operativo Pluri-Regional de Crecimiento Inteligente.
Centro para el desarrollo tecnológico	Proyectos CDTI-ERA-NET	Las ERA-NETs son redes europeas de agencias públicas dedicadas a la financiación de la I+D+i a nivel nacional/regional, que cuentan con el apoyo de la Comisión Europea y cuyo objetivo es favorecer la

Organismo	Programa/línea de financiación	Descripción
industrial (CDTI)		coordinación de los programas de investigación y desarrollo de los estados europeos y movilizar recursos, para afrontar conjuntamente los retos tecnológicos estratégicos de manera más focalizada, coherente y efectiva. La Comisión Europea lanzó el esquema ERA-NET en el sexto programa marco y desde entonces ha ido evolucionando hasta el nuevo instrumento COFUND bajo el programa Horizonte 2020, dotado de una relevante partida presupuestaria y convertido en una herramienta clave para impulsar la investigación, la transferencia de conocimiento y la cooperación internacional, hacia la construcción del Espacio Europeo de Investigación (ERA). Las ERA-NETs ofrecen oportunidades para la internacionalización de la I+D con menores riesgos y mayor tasa de éxito, gracias a la descentralización de la financiación que proveen, que las convierte en acciones puente, a medio camino entre las convocatorias nacionales y las convocatorias europeas. Gracias a estas iniciativas, las entidades españolas pueden participar en proyectos trasnacionales incentivados con financiación pública, mediante procedimientos más accesibles, dado que una vez aprobada una propuesta internacionalmente, se gestiona mediante los programas y fondos nacionales/regionales habitualmente conocidos.
Centro para el desarrollo tecnológico industrial (CDTI)	Proyectos CDTI-Eurostars	La participación de las empresas españolas en los proyectos aprobados en el Programa EUROSTARS-2 se financia con cargo a los Presupuestos Generales del Estado (75%) y a la contribución de la Unión Europea (25%) para el programa. Este apoyo financiero se realiza a través de convocatorias de asignación directa gestionadas por el CDTI denominadas INTEREMPRESAS INTERNACIONAL, a las que solo pueden concurrir las empresas que desarrollen los proyectos de investigación y desarrollo previamente aprobados en las convocatorias del Programa Eurostars-2, siempre y cuando la Secretaría EUREKA les haya comunicado expresamente que cuentan con financiación nacional asegurada mediante subvención.
Unión Europea	Programa Marco de la Unión Europea Horizonte 2020	La Unión Europea (UE) concentra gran parte de sus actividades de investigación e innovación en el Programa Marco de I+D+I, que impulsa el CDTI. Dentro del programa, con el que se busca fomentar la investigación y el desarrollo, la Unión Europea cuenta con un apartado específico destinado a aquellas pequeñas y medianas empresas que, por el riesgo inherente a sus actividades de innovación no tengan acceso a las fuentes de financiación del mercado. Entre sus instrumentos se encuentra el programa Eurostarts.
Generalitat de Catalunya ACCIÓ (Agència per la Competitivitat de l'Empresa)	Comunitats RIS3CAT	Estas ayudas están cofinanciadas con los fondos FEDER, destinados a incentivar la realización de proyectos de I+D+i llevados a cabo en Cataluña y con un impacto en el territorio y en las empresas. Los proyectos han de tener impacto en la internacionalización de los resultados y la tecnología. Los planes de actuación se dividen en: • Innoapat: comunidad para una cadena alimentaria sana, segura y sostenible • Nexthealt: comunidad de soluciones multidisciplinarias para los próximos retos de la salud • Energia: comunidad de energía. • Mobilitat Eco: comunidad de movilidad de bajas emisiones. • Tec-Salut: comunidad para una tecnología aplicada a la salud.

Organismo	Programa/línea de financiación	Descripción
IVF Institut valenciá de finances	Línea IVF innovación + (Horizonte pyme)	El objetivo de esta línea de financiación es impulsar el crecimiento de pymes innovadoras que tengan un proyecto próximo a la fase de comercialización y cuyo objetivo sea dar un fuerte impulso a las actividades necesarias para salir al mercado. Los beneficiarios son las empresas con sede o actividad principal en la Comunidad Valenciana con un proyecto innovador que cuente con una validación técnica suficiente. En este sentido, se considerará que un proyecto esté validado a estos efectos si se ha presentado a la fase II del Instrumento pyme del programa horizonte 2020, y haya superado la puntuación mínima para su aprobación tras la evaluación de la Comisión Europea.

Fondos privados

Organismo	Programa/línea de financiación	Descripción
COFIDES	Programa pyme-invierte	COFIDES E ICEX España Expansión e inversiones han puesto en marcha este programa para ofrecer apoyo integral a la inversión en el exterior de las pequeñas y medianas empresas con el fin de mejorar su competitividad y cubrir sus necesidades globales de implantación en terceros países.
Banco Popular y Comunidad de Madrid	Línea de financiación BEI Comunidad de Madrid	La línea de financiación BEI pymes-midcaps Comunidad de Madrid, está suscrita entre el Banco Popular y la Comunidad Autónoma de Madrid, tiene como finalidad facilitar la financiación de proyectos empresariales realizados en Madrid por este tipo de empresas y autónomos. Entre los proyectos de inversión financiables intangibles se encuentran los gastos en I+D, costes de desarrollo, creación o adquisición de redes de distribución en mercados nacionales o extranjeros dentro de la UE.

Necesidades sectoriales de desarrollo de productos y servicios con mayor potencial

Para completar esta guía de financiación e inversión se ha realizado un análisis sectorial de las oportunidades de inversión a partir de las necesidades sectoriales de desarrollo de productos y servicios con mayor potencial.

- En primer lugar, la oportunidad se encuentra en todas las actividades económicas y administrativas que puedan integrar las nuevas tecnologías, especialmente en **aquellos sectores que requieran una interacción con el usuario final**, dado que los usuarios demandan cada vez más una interacción más natural e inmediata.

A este respecto, destacan las tecnologías de procesamiento del habla, sistemas conversacionales y chatbots que permiten a las empresas que las incorporan ofrecer un servicio de conversación de texto con los usuarios, lo que conlleva un aumento de su disponibilidad y, a su vez, un ahorro de costos relacionado con call centers de atención al cliente.

En esta línea, cabría destacar **el sector de turismo**, intensivo en procesos de cara al público a través de diferentes canales (cara a cara, por teléfono, medios electrónicos, etc), por lo que el lenguaje como interfaz es crítico para las actividades de este sector, en todas las etapas del viaje: antes (agencias de viajes, grandes plataformas de reservas, comparadores de precios, alojamientos, grandes aerolíneas), durante (agencias de viaje, servicio hotelero) y después (valoración de los servicios recibidos, problemas durante el viaje, quejas, etc.).

Por otra parte, destaca el **sector bancario** debido, por un lado, a la cantidad de documentos que tiene digitalizados (hipotecas, normativas internas…) que les interesa analizar, y por otro, motivados por su sección de atención al cliente, tanto por manejar los call centers, como por analizar los motivos por los que se pone el cliente en contacto con el banco, si expresa quejas, etc. Por tanto, la biometría aplicada al sistema financiero es un sector con gran potencial, ya que los bancos están demandando las tecnologías del lenguaje para mejorar sus formas de interaccionar con los clientes.

- En segundo lugar, destaca el **sector Big Data** en una economía mundial que se encuentra en un proceso de transformación digital marcado por la necesidad de gestión de la ingente cantidad de datos y contenidos a través de Internet, redes sociales y medios digitalizados. En este sentido, el procesamiento de esa información y su puesta en valor depende en gran medida del análisis del lenguaje natural. Existe una oportunidad estratégica en la aplicación de las tecnologías del lenguaje natural a los procesos de lenguaje no estructurado, una minería de datos de valor.
- En tercer lugar, destaca el **sector sanitario**, concretamente la **investigación biomédica**, donde las tecnologías del lenguaje tienen tres aplicaciones detectadas como potenciales:
    - ✓ **El soporte a la decisión clínica**: se utiliza la información de los textos, se normaliza y se computa en base a herramientas justificadas que aplican reglas de decisión en base a las guías clínicas a través del procesamiento del lenguaje natural.
    - ✓ **Enriquecimiento de colecciones de datos para la investigación**: la digitalización de la colección de datos que manejan los médicos permite buscar patrones de relación entre cohortes de pacientes.

- ✓ **La codificación de diagnósticos**: es la codificación de información relevante que se necesita para gestionar los hospitales, en base a unos códigos internacionales normalizados, por ejemplo, la clasificación internacional de enfermedades.

- En cuarto lugar, a través del análisis de textos, el análisis de sentimiento y la evaluación de perfiles se pueden **extraer líneas de opinión en relación a un servicio o un producto**, lo que se puede aplicar en diferentes actividades, como la evaluación de currículums, la evaluación de un sistema de consejos de inversión en bolsa, la predicción de resultados de elecciones, etc.
  En esta línea, **la generación automática de textos** permite generar automáticamente informes de eventos deportivos, reseñas periodísticas a partir de datos, informes anuales de facturación de la empresa, etc.

- En quinto lugar, destaca la utilización del procesamiento del habla y la traducción automática en la aplicación de las **"Smart cities"**.

- En sexto lugar, la traducción automática puede ser aplicada en el **sector de enseñanza de idiomas** para apoyar la utilización de medios digitales en la enseñanza.

- Para terminar, el procesamiento del habla y los sistemas conversacionales destacan en la **atención a personas mayores y personas con necesidades especiales**.

Índice de figuras

Índice de tablas

Referencias

[1] D. Pérez y J. d. D. Llorens, «Tecnologías del lenguaje, Plan de impulso y Compra Pública de Innovación,» Noviembre 2017. [En línea]. Available: https://www.astic.es/sites/default/files/articulosboletic/boletic_81-monografico10-david_perez-juan_de_dios.pdf.

[2] M. Palomar, «Tecnologías del Lenguaje Humano aplicadas al aprendizaje de segundas lenguas,» [En línea]. Available: http://www.artic.ua.es/sites/u38/sitio171/PresentacionEducacion.pdf.

[3] B. Núria y R. German, «Informe sobre el estado de tecnologías del lenguaje en España dentro de la Agenda Digital para España,» SESIAD, Madrid, 2015.

[4] J. Listerri, «Tecnologías lingüísticas y sociedad de la información. Economía Industrial (La Sociedad de la Información en España I), 325, 37-56,» 1999. [En línea]. Available: http://liceu.uab.cat/~joaquim/publicacions/Llisterri_99_TecnolLing_SocInfo.pdf.

[5] S. Berner, «"Lost in translation": Cross-Lingual comunication, and virtual academic comunities,» 2015. [En línea]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.2973&rep=rep1&type=pdf.

[6] European Commision, «Status and potential ot the European Language Technology Markets,» 30 January 2014. [En línea]. Available: https://ec.europa.eu/digital-single-market/en/news/lt2013-status-and-potential-european-language-technology-markets.

[7] Cracking the language barrier federation, «Language as a Data Type and key challenge for Big Data,» Julio 2016. [En línea]. Available: http://www.cracking-the-language-barrier.eu/wp-content/uploads/SRIA-V0.9-final-online.pdf.

[8] Kristina, «Kristina project,» [En línea]. Available: http://www.kristina-project.eu/en/.

[9] European Commission, «Europe's Digital Progress Report 2017,» 2017. [En línea]. Available: https://ec.europa.eu/digital-single-market/en/scoreboard/spain.

[10] European Commision , «Cordis,» [En línea]. Available: https://cordis.europa.eu/project/rcn/203292_es.html.

[11] European Commision, «Cordis,» [En línea]. Available: https://cordis.europa.eu/result/rcn/188591_es.html.

Glosario de siglas y acrónimos

AERFAI	Asociación Española de Reconocimiento de Formas y Análisis de Imágenes
AESLA	Asociación Española de Lingüística Aplicada
AETER	Asociación Española de Terminología
BSD	Berkeley Software Distribution
CEEIA	Centro Europeo de Empresas Innovadoras de Valencia
CNAE	Clasificación nacional de actividades económicas
ELRA	European Language Resources Association
EPL	Eclipse Public License
GAIA	Asociación Gaia para la conservación y gestión de la biodiversidad
Langune	Asociación empresarial del sector de Industrias de la Lengua de Euskal Herria
MIT	Massachusetts Institute of Technology
MPL	Mozilla Public License
ONTSI	Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información
RAE	Real Academia Española
REALITER	Red panlatina de terminología y neología de las lenguas romances
RITERM	Red Iberoamericana de Terminología
RSTDA	The Rivers State Tourism Development Agency
RTTH	Red Temática en Tecnologías del Habla
SESIAD	Secretaría de Estado para la Sociedad de la Información y la Agenda Digital
SEPLN	Sociedad Española para el Procesamiento del Lenguaje Natural