

Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España



JULIO 2015

Participantes en la elaboración de este documento:

1. **Editores:** Núria Bel, German Rigau

2. **Panel de expertos que ha redactado el informe:**

- Antonio Colino, Javier Pérez de Vargas, Verónica Verdes-Montenegro (Real Academia de Ingeniería)
- Arantxa Díaz de Ilarraza, Eneko Agirre (Grupo IXA, Universidad del País Vasco)
- Jordi Porta (Centro de Estudios de la Real Academia Española)
- Luis Ignacio Vicente del Olmo (Telefónica I+D) y Juan Rodríguez Vega, Antonio López-Carrasco Comajuncosas y Eduardo Fullea Carrera de la Oficina de Patentes de Telefónica y Estanislao Fernández Gonzalez-Colaço del área de Return on Innovation (Telefónica).
- Luis Navarro Buitrago, María José Chapado, Jaime  **indra** Recio
- Fernando Alana (LANGUNE)
- Ignasi Navarro (INCYTA Multilanguage, SL, representando a Clusterlingua)
- Salvador Estevan Martínez (Dirección de Tecnologías de la Información y la Comunicación de la Administración -- DTIC Jefe de Área Subdirección TIC)
- Salvador Soriano Maldonado (Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información – Coordinador área RISP)
- Fernando de Pablo (Comisión para la reforma de las Administraciones Públicas (CORA))
- Jaime Denis (Asesor, Secretaría de Estado de Investigación, Desarrollo e Innovación)
- M^a Araceli García Martín (Directora de la Biblioteca Agencia Española de Cooperación Internacional para el Desarrollo (AECID))
- Carlos Romero Dexeus (Director de Investigación, Desarrollo e Innovación Turística)
- Joaquín González Gigosos (DG de Formación Profesional, SG de Aprendizaje a lo largo de la vida)
-

Contents

1. Resumen Ejecutivo	7
2. Borrador Análisis DAFO de la situación actual del PLN y TA en España	11
3. Borrador de recomendaciones - Propuestas de actuación	17
4. Introducción	20
5. Recursos lingüísticos en español y lenguas co-oficiales	27
5.1. Estado del procesamiento lingüístico y traducción automática en español y lenguas co-oficiales	27
5.2. Estándares, modelo de licencias, herramientas comunes	35
5.3. Dominios especializados	48
5.4. Asociaciones terminológicas	48
6. Situación de la Investigación del procesamiento lingüístico en España e Iberoamérica	58
6.1. Censo de instituciones y grupos de investigación en procesamiento de PLN y TA	58
6.2. Principales fuentes de financiación de la investigación española en PLN y TA. Ayudas públicas y privadas en España y Europa	66
6.3. Análisis DAFO	74
7. Situación del Mercado y la Industria del lenguaje en español y lenguas co-oficiales	78
7.1. Estudio del sector industrial del PLN en español y lenguas cooficiales	78
7.2. Transferencia de conocimiento sector investigador-industria	87
7.3. Internacionalización	96
7.4. Análisis de la participación de industria TIC	107
8. Las Administraciones Públicas y el PLN y TA en los servicios públicos	127
8.1. Actuaciones en curso, necesidades detectadas y proyectos faro	128
8.2. Proyectos faro de las AA.PP. para impulsar la industria de las tecnologías del lenguaje	136
8.3. Datos lingüísticos abiertos: la AAPP proveedor de datos	138
9. Anexos	142

1.

Resumen Ejecutivo

1. Resumen Ejecutivo

Los sistemas informáticos procesan muy fácilmente datos estructurados. Es decir, información que tiene una estructura y un significado único, explícito y completo. Así, los sistemas informáticos actuales pueden manejar fácilmente bases de datos con millones de registros y datos numéricos. Sin embargo, como el lenguaje natural es la forma más común y versátil que tiene la humanidad de transmitir información, **más de 90% de la información digital disponible es información no estructurada** en forma de textos y documentos (escritos o hablados) en múltiples lenguas¹. Para los sistemas informáticos el problema radica en que toda esta información textual está sujeta a múltiples interpretaciones ya que suele ser incompleta y ambigua.

El procesamiento del lenguaje natural es el camino hacia una comprensión automática cada vez más profunda de la mayor producción humana: el lenguaje.

Para entender la información que contienen estos textos, una capacidad reservada hasta ahora a los humanos, es necesario seleccionar la interpretación correcta y hacer explícita su relación tácita con la realidad: distinguir, por ejemplo, entre 'célula madre', 'célula de silicio', 'célula solar' y en frases como 'los investigadores han demostrado que ciertos tipos de células especializadas se reconvierten fácilmente'.

Las tecnologías de procesamiento del lenguaje natural han desarrollado ya **métodos para hacer explícita la información y para aprender automáticamente cómo relacionarla con la realidad** que describe cada texto para entenderlo y extraer los datos que contiene, traducirlo, resumirlo, etc. A pesar de la dificultad de la tarea, el grado de desarrollo de las tecnologías de Procesamiento del Lenguaje Natural y la Traducción Automática permiten ya muchas de aplicaciones de utilidad, y su rápido desarrollo en la última década augura resultados aún más sorprendentes en un futuro próximo.

Las tecnologías del Procesamiento del Lenguaje Natural (PLN) y Traducción Automática (TA) están en el corazón del software que actualmente procesa la información no estructurada y explota la gran cantidad de datos contenidos en los textos, también los de la web y de los medios sociales. **Las procesadores de PLN y TA son las que hacen posible analizar textos y permitir su explotación**

1

en **aplicaciones informáticas** de uso muy común en sectores tan dispares como la sanidad, el turismo o el mercado de capitales; por ejemplo, la detección de entidades nombradas (nombres propios, fechas, marcas, productos, etc.), la clasificación, agrupación, filtrado y enrutamiento de documentos, la creación de resúmenes automáticos, la extracción de datos, el análisis de sentimientos y minería de opinión, el seguimiento y monitorización de la reputación en los medios sociales, la creación de alertas, la corrección ortográfica y gramatical, la ayuda a la enseñanza de idiomas, la búsqueda inteligente y optimizada de documentos, los sistemas de respuesta automática a preguntas, asistentes personales o la traducción automática de textos. Todas estas aplicaciones se pueden resumir como la **explotación de información no estructurada en documentos y textos** mediante la tecnología de PLN y TA.

El estudio de LT-innovate (2013) calcula un mercado de 26.500M€ en 2015 que puede llegar a 65.000M€ en 2020 (se incluye el mercado de la traducción, más consolidado que el de otras áreas de aplicación de esta tecnología).

Aplicadas de la forma adecuada, estas tecnologías pueden ayudar a las empresas y organizaciones a **optimizar muchos de sus procesos productivos y a obtener información y conocimiento muy valioso** de sus propios datos y de los datos disponibles en un mundo cada vez más digitalizado y global. Las tecnologías del lenguaje constituyen una **industria habilitadora** que participa de forma horizontal en multitud de aplicaciones y dispositivos.

El mercado de estas tecnologías está creciendo rápidamente y los informes de varias consultoras (informes de Gartner, LT-Innovate, AltaPlana, etc.) pronostican un **gran crecimiento del mercado mundial en 2020** basándose en la explosión de aplicaciones observada en los últimos dos años y en un crecimiento exponencial de los datos textuales digitales.

El español es la segunda lengua con más hablantes nativos en el mundo y la tercera por número de hablantes, posición que se mantiene en los rankings de lenguas más usadas en las redes sociales. Además, aunque represente actualmente el 6% de la población mundial (con más 500 millones de hablantes), se prevé que en 2030 sea ya la segunda lengua de intercambio económico en el mundo, principalmente por el crecimiento del mercado latinoamericano, y norteamericano (actualmente con 52 millones de hablantes de español en los EEUU)².

Sin embargo, el desarrollo de aplicaciones para procesar textos escritos en una lengua (o idioma) depende de la disponibilidad de la tecnología para esa lengua. Actualmente, la mayoría de aplicaciones salen al mercado disponibles solamente para procesar el inglés. Para que puedan procesar otros idiomas es necesario disponer de datos sobre cada uno de ellos. Son los llamados recursos lingüísticos: grandes cantidades de textos que sirven de ejemplo para los algoritmos de aprendizaje automático, diccionarios con información de diferente índole en los que basar decisiones, y procesadores lingüísticos especializados en ese idioma.

² http://eldiae.es/wp-content/uploads/2015/06/espanol_lengua-viva_20151.pdf

Por tanto, en el caso de España la prometedora expansión del mercado mencionada antes supone **una oportunidad de negocio en el que las empresas españolas pueden ser líderes de mercado de las aplicaciones en español y de gestión del multilingüismo**, dada su experiencia con cuatro lenguas co-oficiales. Pero la producción de recursos lingüísticos para cada lengua y dominio de aplicación tiene, todavía, un coste elevado que no pueden asumir las pequeñas y medianas empresas que podrían aprovechar esta oportunidad. **Para garantizar la creación de aplicaciones en español y las demás lenguas co-oficiales hay que aumentar el número, la calidad, la variedad y la disponibilidad de los recursos y procesadores que les dan soporte como una infraestructura común y reducir así la inversión inicial que permitiría a las pymes españolas aprovecharse de esta oportunidad.**

En España **existe un número creciente de pequeñas y medianas empresas que desarrollan productos en español, catalán, euskera o gallego para clientes o nichos específicos**: gestión de la atención al cliente, análisis de reputación de hoteles y empresas, monitorización de la opinión de clientes en medios sociales, análisis de informes clínicos, traducción automática de noticias, sistemas de ayuda a la traducción, etc.



Un obstáculo importante para el pleno desarrollo de esta industria es la poca visibilidad que las tecnologías de PLN y TA tienen en el tejido industrial español de las TIC. Los informes de Gartner, entre otros, alertan del desconocimiento que existe sobre las aplicaciones que las utilizan. La falta de demanda de productos mantiene a las empresas españolas del sector en unas dimensiones reducidas (de entre 1 y 10 trabajadores, con alguna excepción), sin conciencia sectorial ni una estructura de representación que agrupe sus intereses. Para maximizar el beneficio de oportunidad de negocio de las empresas españolas, debería reforzarse a las pymes españolas con programas de apoyo para la mejor comercialización e internacionalización de sus productos.

La administración y las grandes empresas del sector TIC desconocen en buena medida las potencialidades de las tecnologías de PLN y TA. Además, tienen dificultades para evaluar el retorno de una inversión en innovación en esta área y algunos grupos de investigación y desarrollo en grandes empresas españolas, activos en este sector hace 15 años, han desaparecido. En la academia, con más de 30 grupos de investigación, la falta de crecimiento del sector causa que la mayor parte de doctores formados en España se vean obligados a buscar oportunidades en multinacionales extranjeras, **con la consiguiente pérdida de inversión en formación de especialistas altamente cualificados.**

Dadas las múltiples aplicaciones donde el PLN y la TA pueden resultar críticas para optimizar procesos y recursos productivos en múltiples ámbitos empresariales y de las administraciones públicas, **es urgente que la formación en estas tecnologías llegue a los profesionales de las TIC en empresas y administración pública.**

Para maximizar la capacidad emprendedora y la reutilización, procesadores y datos comunes deberían ofrecerse, como ya se hace en otros ámbitos, en una infraestructura de apoyo a la innovación que permita la creación rápida de prototipos y su fructífera comercialización.

Por otro lado, aumentar **la demanda de esta tecnología en la administración y las grandes empresas puede actuar de catalizador y tractor** que acelere el pleno desarrollo del sector. Su implicación daría visibilidad, prestigio y garantía a las aplicaciones desarrolladas dinamizando de una manera clara el desarrollo del mercado en España y también en los países de habla hispana que pueden beneficiarse de los productos en español desarrollados por nuestra industria.

Además, la participación de las administraciones permitiría también que los recursos y procesadores del español (y lenguas co-oficiales) pudieran ser ofrecidos como **datos abiertos** con los que fomentar la innovación basada en la reutilización de datos públicos, uno de los objetivos de la Agenda Digital para España y de la Agenda Digital para Europa.

En Europa se están llevando a cabo iniciativas similares, y la gestión del multilingüismo en el mercado digital único europeo, está incluido en los objetivos prioritarios del programa *Connecting Europe Facility* (CEF) con la **puesta en marcha de servicios públicos digitales a nivel europeo**, incluyendo servicios lingüísticos y mecanismos de acceso a recursos lingüísticos disponibles para las lenguas de Europa.



2.

Borrador Análisis DAFO
de la situación actual del
PLN y TA en España

2. Borrador Análisis DAFO de la situación actual del PLN y TA en España

Análisis DAFO de la situación actual del PLN y TA en España

En el presente informe se ha llevado a cabo un estudio sobre la situación actual del PLN y TA en España, recopilando información sobre datos objetivos que sustentan un análisis en términos de las debilidades, amenazas, fortalezas y oportunidades del sector. Este análisis DAFO sustenta un conjunto de recomendaciones sobre las acciones necesarias para mejorar el crecimiento óptimo, la competitividad y la internacionalización de las empresas españolas del sector.

Oportunidades

- El procesamiento del lenguaje natural y la traducción automática son consideradas por las grandes consultoras una **tecnología clave** para el análisis de Big Data no estructurado.
- La **previsión de una creciente explosión de aplicaciones en múltiples áreas**, por ejemplo, Turismo, Sanidad, Justicia, Educación, etc. en las que las tecnologías del procesamiento del lenguaje natural pueden optimizar y sistematizar procesos horizontales.
- La **demanda urgente de aplicaciones** relacionadas con los medios sociales, Big Data y datos abiertos proporciona objetivos de desarrollo a corto plazo que ya pueden ser satisfechos en consorcios mixtos (universidad-empresa-administración pública) que maximizarían su visibilidad y contribuirían al desarrollo del mercado y la industria.
- La existencia de un amplio mercado potencial para productos en español. Además del gran mercado latino-americano, las estadísticas confirman el

creciente **aumento de hispanohablantes en EEUU**, llegando a 52 millones en 2015³, con alto poder adquisitivo.

- Hay más de **300 investigadores en Sudamérica**, la mayoría en Méjico, con muy buenas relaciones con universidades de EEUU y pueden convertirse en colaboradores de empresas españolas.
- El **momentum en Europa del PLN y TA**, donde el multilingüismo está siendo potenciado gracias a estas tecnologías. Se están desarrollando plataformas e infraestructuras para el procesamiento lingüístico multilingüe, y hay asociaciones industriales, fundaciones y distribuidores que han trabajado ya de forma colaborativa los aspectos formales: estándares, licencias, modelos de negocio, etc.
- También existen **modelos europeos e internacionales** de oferta de recursos lingüísticos en portales de datos abiertos a los que sumarse.

Fortalezas

- **Capacidad y experiencia en la producción de recursos y procesadores para hacer el procesamiento básico de PLN y TA** para el castellano, catalán, vasco y gallego, además del inglés.
- La existencia de **más de 30 grupos de investigación universitarios**, consolidados y organizados en asociaciones y redes, que han creado 9 spin-offs. Sus líneas de investigación en PLN abarcan casi todos los ámbitos en los que se trabaja actualmente a nivel internacional, y no solamente para el español y lenguas co-oficiales (ver cap.6).
- Los **investigadores españoles** participan en proyectos, asociaciones y grupos de estandarización europeos e internacionales.
- Los grupos de investigación tienen **probada experiencia en transferencia de tecnología** gracias a su participación en proyectos de I+D+i europeos.
- Las **academias y organismos especializados** en España han desarrollado y mantenido materiales de español y lenguas co-oficiales que pueden ser utilizados rápidamente como recursos lingüísticos de lenguas de especialidad.
- Existen de **bases de datos terminológicas**, de sinónimos, thesaurus de topónimos, etc. exportables globalmente.
- Las **instituciones relacionadas con las lenguas en España** han ejercido y siguen ejerciendo una labor reguladora y estandarizadora sobre la lengua.
- La RAE tiene una **dimensión panhispánica**, forma parte de redes con organismos equivalentes en latinoamérica, y gozan de prestigio en la sociedad.
- La **experiencia de acciones estratégicas** de apoyo a la I+D+i en los planes nacionales ha demostrado su eficacia.

³ http://eldiae.es/wp-content/uploads/2015/06/espanol_lengua-viva_20151.pdf

- La existencia de un **tejido industrial en TA** con experiencia que tiene como clientes prioritarios las AAPP.
- Las empresas españolas del sector tienen **gran experiencia en gestión del multilingüismo**, lo que puede ser un modelo exportable.
- Gracias a los programas nacionales de transferencia (CENIT, CIEN, Doctorados industriales) los grupos de investigación con experiencia en transferencia en programas europeos pueden rápidamente establecer **modelos de colaboración**.
- Cuatro empresas españolas del sector han **abierto oficinas en EEUU y Sudamérica** demostrando la viabilidad de la internacionalización de nuestra tecnología.
- La existencia de un sector TIC español potente que está demostrando su capacidad para competir globalmente.
- España es el **mejor líder tecnológico del mercado potencial de aplicaciones en español**, con un mercado potencial con más de 500M de hablantes.
- La Agenda Digital para España (Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información) ha definido objetivos a los que puede contribuir el sector del PLN y TA.
- **Gran conocimiento del negocio** de sectores específicos: salud, seguridad, justicia, educación, turismo, hacienda, etc. y sus necesidades prioritarias y masivas.
- La **adhesión de España a la directiva europea** de reutilización de datos de las AAPP para la creación de negocio y la puesta en marcha de portales de datos abiertos en todos los organismos públicos (incluidas las bibliotecas).
- La existencia de tecnología española (en la academia y la industria) para desarrollar **recursos y procesadores para el PLN y TA** para el español y las lenguas co-oficiales al mismo nivel de calidad de las equivalentes disponibles para inglés.

Debilidades

- **Desconocimiento de las características de esta tecnología** de alto coste de inversión en localización y adaptación (a idiomas y dominios de aplicación) con recursos y procesadores lingüísticos específicos.
- **Desconocimiento del área de PLN y TA** y de la utilidad de los datos lingüísticos (textos, glosarios, ontologías, tesauros, etc.) generados por las AAPP.
- Los recursos y procesadores, estándares de interoperabilidad, licencias y modelos de negocio ya consensuados en Europa todavía **no son ampliamente conocidos por las empresas en España**.
- La mayoría de empresas del sector son pymes y microempresas con escasa **capacidad de inversión** por falta de capitalización (la media de facturación anual es de 1,6M€) y que gastan en el desarrollo de recursos y procesadores

lingüísticos propios, ya que desconocen los existentes o no pueden financiar su compra.

- Las empresas españolas, con una media de 10,5 empleados solo tienen una media 1 doctor especialista empleado, mientras que la media europea es de una contratación de doctores del 30%.
- La mayoría de empresas del sector son pymes y microempresas con **poca capacidad de comercialización e internacionalización** de sus productos.
- Falta de una **asociación empresarial** que actúe como representante de los intereses industriales del sector en España y proponga estrategias y busque apoyo especialmente en la comercialización de productos y la internacionalización de las empresas españolas.
- El número y cobertura de recursos y procesadores de PLN y TA para el castellano, catalán, vasco y gallego es aún bastante **inferior a los disponibles para el inglés**.
- La falta de una infraestructura común y asequible imposibilita que el sector cree *startups*.
- Carencia de recursos y procesadores lingüísticos para el **español hispanoamericano**.
- La **poca colaboración** entre empresas y entre empresas y grupos de investigación impide la reutilización de datos y procesadores lingüísticos, elevando la inversión que deben realizar las empresas, restándoles efectividad en otros ámbitos como la innovación, la promoción comercial y la internacionalización.
- A pesar del esfuerzo de las academias españolas por crear materiales lingüísticos que podrían ser cruciales para el PLN y TA, la **falta de conocimiento e inversión coordinada** dificulta la implantación de modelos de disponibilidad y licencias que garanticen la re-utilización.
- Ofrecer **datos lingüísticos de calidad** en portales de datos abiertos **requiere de inversión y planificación** convirtiéndose en una actividad marginal de poco interés y poco sostenible sin financiación específica.
- El PLN y la TA **no están reconocidos como una disciplina en la academia nacional**. Es un ámbito de investigación interdisciplinar y la formación en estas tecnologías está dispersa por diferentes especialidades lo que dificulta su visibilidad.
- Los grupos de investigación españoles han establecido más **relaciones con empresas europeas**, gracias a la financiación europea de I+D, que con la industria española con poca capacidad para invertir en innovación.
- La **financiación de la investigación y la innovación** en PLN y TA es discontinua y dificulta el progreso y el mantenimiento de equipos de trabajo especializados. Esta situación tiene un impacto negativo en los grupos de investigación nacionales que se sobredimensionan para aprovechar la financiación puntual pero no pueden mantener los equipos ya formados.

- Hay también una falta general de coordinación de los programas de financiación españoles con los europeos.
- Pérdida de **inversión en formación** de especialistas altamente cualificados por falta de un mercado laboral estable en el sector.
- **Desconocimiento de las capacidades de esta tecnología** para lograr procesos mucho más eficientes en múltiples áreas de aplicación en la administración, que si fuera usuaria crearía además referencias que apoyarían la comercialización y la internacionalización de las empresas.

Amenazas

- **Desarrollo en aumento del sector en Europa y EUA que puede entrar en el mercado de aplicaciones en español.**
- Primacía actual del **mercado norteamericano**, también para el español, que lidera el desarrollo tecnológico en la búsqueda de tecnologías independientes de la lengua.
- Cada vez es más común la **comercialización de las aplicaciones a través de APIs** en forma de servicios que pueden ser prestados desde la nube de forma deslocalizada.
- Diferentes **asociaciones y organizaciones europeas** han consensuado ya estándares y modelos de licencias de forma colaborativa lo que puede arrinconar a la industria española si no participa más activamente.
- El uso del español, y lenguas co-oficiales, como lenguas de dominios especializados **desaparecerá** si no se fomentan las publicaciones en español y su disponibilidad y uso en el mundo digital.
- El coste del desarrollo de recursos para diferentes idiomas propicia que la investigación se centre en conseguir tecnología independiente de la lengua, afectando negativamente la especificidad de un potencial mercado del español.



3.

Borrador de
recomendaciones -
Propuestas de actuación

3. Borrador de recomendaciones - Propuestas de actuación

El presente informe ofrece un estudio sobre el sector de la industria del Procesamiento del Lenguaje Natural y la Traducción Automática en España. En su elaboración han participado representantes de empresas del sector, en su gran mayoría pymes agrupadas en algunos casos en clusters, grandes empresas del sector TIC, representantes de la administración pública y de grupos de investigación españoles.

El estudio constata que **existe la masa crítica** en el sector del PLN y TA -- constituida por empresas, instituciones y grupos de investigación-- **y una incipiente demanda de aplicaciones en todos los sectores de actividad económica, y en especial los de la administración pública**. Estos dos son los elementos potencialmente suficientes y necesarios para que se constituya una próspera industria española en un sector con futuro para el que todas las consultoras internacionales pronostican un **gran crecimiento del mercado mundial de aquí al 2020**. Una industria de las consideradas industrias del futuro, que pueda mejorar y ofrecer nuevos servicios digitales y productos de uso habitual para todos los ciudadanos.

El estudio ha observado no obstante que, en España, el sector sufre un estancamiento debido a diferentes factores y que solamente una serie coordinada de medidas estratégicas puede catalizar y desbloquear la actual situación además de acelerar su pleno crecimiento. Es importante actuar rápidamente ya que, como también se desprende del análisis realizado, se puede aprovechar el *momentum* actual para que algunas actuaciones maximicen su impacto: está prevista una explosión de aplicaciones en los próximos cinco años, pero son aplicaciones que necesitan componentes específicos para cada idioma y dominio de aplicación, como se constata en estudios realizados en Europa, aquellas lenguas que no dispongan de ellos quedarán excluidas. Si la demanda no se puede satisfacer en un tiempo razonable, se extinguirá. Por tanto es crucial que haya productos que cubran la demanda del mercado nacional, para español y lenguas coficiales, y, en el caso del español, la demanda de un gran mercado internacional. También se desprende de los datos que este gran mercado internacional de las aplicaciones para español puede ser cubierto por empresas de otros países si las desarrollan antes.

Una serie de actuaciones estratégicas llevadas a cabo en el momento adecuado debe servir para visibilizar este sector y dinamizar su crecimiento, asegurándose de que éste sea sostenido.

El estudio sugiere que las actuaciones habrían de hacer **énfasis en reunir y estructurar adecuadamente a todos los agentes implicados** ajustando sus respectivos roles en la cadena de I+D+i para crear un crecimiento óptimo y sostenible. Con las medidas adecuadas, los grupos de investigación en PLN y TA y las empresas especializadas sustentarán la **Investigación**. La administración pública y las grandes TIC darán el impulso inicial al **Desarrollo** de aplicaciones que las popularicen y las empresas de la industria del PLN y TA podrán concentrar sus recursos, limitados, en la **Innovación**, la comercialización y la internacionalización.

El estudio realizado sugiere, además, acciones cuyo objetivo es **optimizar la cadena de valor** de la industria de las tecnologías del PLN y TA, reduciendo los costes de inversión en recursos y procesadores para lenguas específicas mediante la creación de una infraestructura lingüística común financiada con fondos públicos. Al reducir la inversión, permitirá que pymes ya creadas o futuras *startups* puedan concentrar su actividad en el desarrollo de aplicaciones innovadoras que cubran el gran número de sectores que pueden beneficiarse de ellas.

Por último, este estudio también ha detectado que las empresas del sector reclaman medidas de apoyo para la comercialización e internacionalización de sus productos conscientes del gran mercado internacional al que no pueden llegar por su limitada capacidad financiera.

En particular, el estudio sugiere las siguientes medidas:

- 1) Aumentar el número, la calidad, la variedad y la disponibilidad de los recursos y procesadores lingüísticos que dan soporte a la utilización de esta tecnología para el español y las lenguas co-oficiales para garantizar la misma disponibilidad de aplicaciones avanzadas que para el inglés.** Actualmente, hay una primacía de la industria de PLN y TA global en inglés y en particular en el mercado norteamericano, donde el uso del español está aumentando y pronto será considerado como objetivo.
- 2) Estructurar y fomentar la colaboración entre los agentes (administración, industria y los grupos de investigación).** La poca colaboración entre empresas y entre empresas y grupos de investigación limita la reutilización de datos y procesadores comunes y aumenta innecesariamente las inversiones que realizan las empresas en desarrollar sus propios procesadores y recursos lingüísticos, restándoles efectividad en otros puntos de la cadena de valor como la innovación, la promoción comercial y su internacionalización. **Los interlocutores pueden ser la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información (SETSI) por parte de la administración, la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) por parte de los grupos de investigación, y una organización empresarial, inexistente todavía, que represente los intereses del sector de PLN y TA en España.**

- 3) **Hacer visibles las tecnologías de PLN y TA** en la administración y las grandes empresas del sector TIC mediante acciones de difusión y formación para profesionales TIC.
- 4) **Impulsar a las empresas españolas a consolidar una oferta de productos de calidad para el español y mejorar su comercialización para llegar a tiempo de liderar el mercado internacional** de productos para el español. La contratación pública y la compra pública innovadora pueden hacer de tractor del mercado en el sector TIC para hacer visible su potencialidad en demostradores que aumenten la confianza de la demanda del mercado nacional e internacional. También es la contratación de la administración pública y las buenas relaciones público-privadas con las administraciones latinoamericanas (que se podrían fortalecer con convenios bilaterales) las que pueden dinamizar la internacionalización de estos productos y servicios en el menor tiempo posible.
- 5) **Diseñar y poner en marcha laboratorios virtuales que reúnan la infraestructura lingüística y tecnológica**, la capacidad de computación de altas prestaciones y la formación necesarias para empresas del sector PLN y TA, TIC y futuras startups. Todas las consultoras internacionales coinciden en que el PLN será una tecnología clave para el análisis de Big Data no estructurado. Se espera una explosión de productos innovadores que ya surgen de startups y spinoffs en Europa y América. Hay que trasladar la probada excelencia investigadora española y sus resultados a la industria que debería ser apoyada para conseguir optimizar la cadena de valor del sector reduciendo la inversión en recursos y procesadores lingüísticos que pueden ser comunes.
- 6) **La experiencia de la administración y las empresas españolas en la gestión de las cuatro lenguas co-oficiales puede hacer de España un punto de referencia** en la gestión del multilingüismo y las soluciones que ofrece la tecnología del PLN y TA. En Europa hay una creciente preocupación por el obstáculo que supone la diversidad lingüística para el Mercado Único Digital Europeo, por ejemplo.
- 7) **Fortalecer y mantener el esfuerzo en la investigación, el desarrollo y la innovación en tecnologías en PLN y TA más allá de la explotación de recursos y procesadores lingüísticos en español o multilingües.** Garantizar una financiación continuada en I+D+i, con programas de retención de talento, doctorados industriales y la inclusión de esta especialidad en programas específicos servirá para asegurar una posición de la industria y la investigación española más allá de la explotación de recursos y procesadores lingüísticos en español o multilingües.

4.

Introducción

4. Introducción

El procesamiento del lenguaje natural es el **camino hacia una comprensión automática cada vez más profunda de la mayor producción humana: el lenguaje**. El lenguaje es la forma más común y versátil de comunicación. Los sistemas informáticos procesan fácilmente datos, es decir, información que tiene una estructura y un significado único y explícito (información estructurada). Pueden manejar fácilmente tablas con millones de datos numéricos, por ejemplo.

La **capacidad de “entender” textos**, reservada hasta ahora al cerebro humano, se basa en la capacidad de seleccionar un significado entre los posibles a partir de los significados posibles de todas las palabras que se combinan en ellos y su relación tácita con la realidad. Las tecnologías de procesamiento del lenguaje natural han logrado ya **métodos para hacer explícita la información y aprender automáticamente a relacionarla con la realidad** que describe cada texto para entenderlo, traducirlo, resumirlo, etc. A pesar de la dificultad, el grado de desarrollo de las tecnologías de procesamiento del lenguaje natural permite hoy multitud de aplicaciones de utilidad, y su rápido desarrollo en la última década augura resultados cada vez más sorprendentes.

Las tecnologías de Procesamiento del Lenguaje Natural (PLN) y Traducción Automática (TA) están en el **corazón del software que actualmente procesa el lenguaje humano y explota la gran cantidad de datos textuales de la web y de los medios sociales**. El desarrollo de estas está considerado en la actualidad como prioritario frente al desarrollo del procesamiento de voz. Las herramientas de PLN y TA son las que hacen posible analizar textos y facilitar su explotación en aplicaciones informáticas de uso muy común en sectores tan dispares como la sanidad, el turismo o el mercado de capitales. Por ejemplo, la detección de entidades nombradas (nombres propios, fechas, marcas, productos, etc.), clasificación, agrupación, filtrado y enrutamiento de documentos, creación de resúmenes automáticos, extracción de información, análisis de sentimientos, minería de opinión, seguimiento y monitorización de la reputación en los medios sociales, creación de alertas, corrección ortográfica y gramatical, enseñanza de idiomas, búsqueda de inteligente y optimizada, sistemas de respuesta automática a preguntas y asistentes personales, la traducción automática de textos, etc. Todas estas aplicaciones se pueden resumir como la **explotación de información no estructurada en documentos y textos** mediante técnicas de PLN y TA.

Más de 90% de la información digital disponible es información no estructurada: textos, producciones del lenguaje humano: textos, escritos o hablados, que se caracterizan por tener más de un significado posible y por contener información no explícita.



El sector de las tecnologías del lenguaje ha sido retratado en profundidad en diversos informes recientes. En la actualidad el sector industrial mundial se compone de unas **2.500 empresas y más de 800 centros de investigación** en procesamiento de lenguaje natural. Según estima el informe "Report on the State of the European Language Technology Industry", de la asociación empresarial europea LT-Innovate⁴, en el año 2013 el mercado mundial de las tecnologías del lenguaje tenía un **volumen aproximado de unos 19.300 M € y se prevé que en 2015 alcance los 30.000 M €**. Según el mismo informe el mercado europeo de interacción por medio de la voz crece a un 9.7% anual y llegará en 2015 a 8.600 M €. El mercado europeo del procesamiento lingüístico de contenidos crecerá hasta los 6.200 M €. El sector de la traducción automática en Europa alcanza hoy en día la cifra de 8.600 M € y **llegará al valor de 14.900 M € en 2015**.

En cuanto a idiomas, **el español es la lengua más hablada del mundo después del chino**, si tenemos en cuenta los hablantes nativos, y tercera por número de hablantes después del inglés. Sus hablantes representan el 6% de la población mundial: 500 millones. Se prevé que en 2030 sea la segunda lengua de intercambio económico en el mundo, principalmente por el crecimiento del mercado latinoamericano⁵.

El uso del mismo idioma aumenta por tres la cuota de exportaciones de España a los países de habla española⁶. **Nueve de cada diez usuarios de esta tecnología se encuentran fuera de nuestras fronteras**. Además de la comunidad latinoamericana existen otras comunidades hispanohablantes en Norteamérica con un enorme potencial de crecimiento.

Otras lenguas cooficiales como el gallego pueden representar una **ventaja competitiva para la hora de desarrollar herramientas** de procesamiento del lenguaje en países donde se habla portugués.

Por otra parte, España cuenta con **prestigiosas organizaciones internacionales especializadas** en la lengua española y las lenguas cooficiales; entre ellas destacan la Real Academia Española y el Instituto de España.

Por otra parte, los investigadores y una buena parte de la Industria del sector se encuentran **agrupadas alrededor de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)**, que ha cumplido su 30 aniversario este año.

Hasta el momento el mercado de la traducción automática domina el mercado europeo. Se trata de un sector estratégico para la Unión Europea dado que cuenta con 24 lenguas oficiales, algunas de ellas en peligro de extinción digital, como señala el estudio de la Red de Excelencia META-NET "Europe's languages in the Digital Age" .



4

<http://www.lt-innovate.eu/lt-observe/document/lt-innovate-report-state-european-language-technology-industry>

⁵ http://eldiae.es/wp-content/uploads/2015/06/espanol_lengua-viva_20151.pdf

⁶ CARRERA TROYANO, M. y OGONOWSKI, Michal. El valor económico del español: España ante el espejo de Polonia. Madrid: Instituto Complutense de Estudios Internacionales, 2008

En España existen unos **30 grupos investigadores de reconocido prestigio internacional** trabajando en procesamiento del lenguaje español y lenguas cooficiales así como grupos especializados en semántica. Asimismo se han configurado algunas agrupaciones empresariales en Cataluña, Euskadi y Madrid alrededor de la industria del lenguaje (ver cap.6).

Mientras que existe el potencial para disponer de una industria de éxito, en el estudio realizado por la red de excelencia META-NET, se señala que la **carencia de recursos para el procesamiento del lenguaje natural** en una lengua particular es la barrera más importante para el desarrollo en Europa de aplicaciones de la industria del lenguaje, como son:

- **Optimización de procesos industriales** de gestión lingüística de documentación: traducción de documentos y Herramientas de autor (correctores, generación de documento, etc.)
- **Comunicación y asistencia personal** (asistentes virtuales, comunicación hombre-máquina para coches, atención al cliente e interacción con robots; buscadores inteligentes y respuesta automática de preguntas).
- **Procesamiento inteligente de información y conocimiento** (extracción y minería de información de textos y contenidos, clasificación de documentos, resumen automático, análisis forense de documentos, etc.)
- **Aprendizaje de lenguas y aprendizaje mediatizado** por ordenador.
- Tecnologías para **contenidos creativos**.

Este estudio de META-NET identifica los recursos lingüísticos, esto es los componentes y conjuntos de datos que estas aplicaciones necesitan para cubrir el mercado europeo, y alerta claramente sobre la situación actual en la que hay importantes carencias en cuanto a la disponibilidad de estos datos para cada una de las lenguas europeas comparándolas con la disponibilidad para el inglés, que domina claramente el mercado. El estudio se basó en la **calificación de siete aspectos básicos**: cantidad de recursos, disponibilidad, calidad, cobertura lingüística, madurez, mantenimiento y adaptabilidad. El resultado, en forma de ranking del nivel de recursos existentes que darían soporte a la creación de aplicaciones **denuncia claramente la situación del español con un nivel significativamente por debajo del disponible para el inglés** a pesar de su situación pareja como lenguas del mundo, y todavía más precario para las lenguas co-oficiales.

La red de excelencia META-NET también publicó en 2013 una **agenda estratégica que identificaba las acciones necesarias para desbloquear el desarrollo de las tecnologías de la lengua** aportando una infraestructura de apoyo en todas las lenguas de Europa. La Comisión de la Unión Europea ya ha adoptado algunas de estas recomendaciones en convocatorias del programa marco H2020⁷ y en la definición del programa *Connecting Europe Facility* (CEF)⁸

La conclusión de este estudio es clara: para garantizar la disponibilidad de aplicaciones en español y lenguas co-oficiales en España hay que aumentar el número, la calidad y la disponibilidad de los recursos que les dan soporte.

⁷<http://ec.europa.eu/programmes/horizon2020/>⁸<http://ec.europa.eu/digital-agenda/en/connecting-europe-facility>

que pretende la puesta en marcha de servicios digitales a nivel paneuropeo, incluyendo servicios lingüísticos y con un apartado específico para garantizar un mecanismo de coordinación de información y acceso a recursos lingüísticos disponibles para las lenguas de Europa.

En los programas europeos, no obstante, el suministro efectivo de recursos que constituye la infraestructura de apoyo para cada lengua no queda cubierta. **Es necesario garantizar el suministro para el español y las lenguas co-oficiales**, ya que, a pesar de que existe una industria de las tecnologías del lenguaje, sus características (PyMEs en su gran mayoría) hacen imposible pensar que ésta se autoabastezca con la rapidez necesaria.

El análisis realizado por estas organizaciones indica que **es necesario apoyar este sector, también en España, con un programa específico de medidas** que incluya la definición de una estrategia de desarrollo para el sector y su implementación en un período temporal corto así como la creación de la infraestructura necesaria para darle soporte efectivo. Ya existen antecedentes en otros países de la Unión.

En el marco de sus competencias, la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información (en adelante SETSI) es **responsable de la ejecución y coordinación de la Agenda Digital para España**. La Agenda Digital para España marca la hoja de ruta en materia de Tecnologías de la Información y las Comunicaciones (TIC) y de Administración Electrónica para el cumplimiento de los objetivos de la Agenda Digital para Europa en 2015 y en 2020. Incorpora grandes objetivos para el desarrollo de la economía y la sociedad digital en España cuyo desarrollo se verá impulsado de forma importante por el uso de tecnologías del lenguaje, entre ellos:

- Desarrollo de la economía digital para el crecimiento, la competitividad y la internacionalización de las empresas españolas
- Mejora de la administración electrónica y los servicios públicos digitales
- Impulso del I+D+i en las industrias de futuro
- Promoción de la inclusión y alfabetización digital y la formación de nuevos profesionales TIC
- Plan de TIC en PYME y comercio electrónico
- Plan de Internacionalización de empresas tecnológicas

Por todo ello, **la SETSI puede incorporar un nuevo Plan de impulso a las tecnologías del lenguaje en la Agenda Digital para España (ADpE)** como un instrumento adecuado para el desarrollo de este sector estratégico.

El citado Plan es una **iniciativa interdepartamental por el carácter multidisciplinar** de las tecnologías del lenguaje, y tiene como objetivo coordinar todas las actuaciones de la Administración General del Estado para el impulso de la industria del lenguaje.

El informe Report on the State of the European Language Technology Industry describe el estado actual del mercado constatando la existencia de más de 500 empresas relacionadas con este sector. Se trata de PyMEs concentradas en el norte de Europa que no pueden abarcar la totalidad del mercado europeo o internacional. El informe recomienda apoyar a estas empresas con una infraestructura específica para disponer de recursos y tecnología que permitan desbloquear la creación de aplicaciones para otras lenguas además del inglés.



Por ello, para la elaboración del Plan de impulso de las tecnologías del lenguaje se ha constituido un **Comité Directivo interdepartamental que reúne a los órganos de la Administración General del Estado con protagonismo en esta materia**. El Comité Directivo, a su vez, ha requerido un informe previo sobre el estado de las tecnologías del lenguaje en España que incluya, además, recomendaciones para el desarrollo de este sector en España.

Presentamos aquí el informe requerido por el Comité Directivo para **el Impulso de la Industria del Lenguaje Natural**. En este informe se ha llevado a cabo, como solicitó el Comité Directivo, un **análisis y diagnóstico inteligente de la situación actual** sobre las áreas de interés de ese comité:

- **Recursos lingüísticos en español y lenguas cooficiales:**
 - Estado del procesamiento lingüístico y la traducción automática (español y lenguas cooficiales).
 - Situación de los procesadores lingüísticos, recursos lingüísticos y semánticos, corpus paralelos, así como colecciones textuales de gran tamaño anotadas con información lingüística de alta calidad. Número, calidad y modelo de licenciamiento.
 - Estándares, modelo de licencias, herramientas comunes.
 - Estándares para recursos y procesadores lingüísticos, modelos de licencias y herramientas comunes para la anotación de corpus.
 - Dominios especializados
 - Ej.: medicina, legislación, ingeniería, etc.
- **Situación de la Investigación del procesamiento lingüístico en España e Iberoamérica:**
 - Censo de instituciones y grupos de investigación en España e Iberoamérica.
 - Instrumentos financieros disponibles y eficacia.
 - Incluye España y UE.
- **Situación del Mercado y la Industria del lenguaje en español y lenguas cooficiales:**
 - Estudio del sector del PLN en español y lenguas cooficiales (España, Iberoamérica, Internacional).
 - Censo de empresas y principales proyectos; análisis de las características de empresas españolas, estudio de mercado internacional actual y futuro. Competitividad internacional.
 - Transferencia de conocimiento sector investigador-industria.
 - Situación actual y casos de éxito; instrumentos disponibles y análisis de eficacia.

Este informe ha sido elaborado con el asesoramiento de un Comité de 12 expertos: personas relevantes en el ámbito de las tecnologías del lenguaje de los sectores de la investigación, académico y la industria y representantes de las administraciones públicas relevantes en los ámbitos de aplicación interesantes para éstas.



- Internacionalización.
- Situación actual y casos de éxito; instrumentos disponibles y análisis de eficacia.
- Análisis de posibilidades de participación de industria TIC en el Plan.
- **Las Administraciones Públicas y el procesamiento de lenguaje natural y la traducción automática en los servicios públicos.**
 - Actuaciones en curso, necesidades detectadas y proyectos faro.
 - Infraestructuras y servicios comunes; ámbitos especializados – verticales-.
 - Datos lingüísticos abiertos (AAPP como fuente de datos).
 - Catálogo de recursos de interés, Modelo de licencia y distribución).

A tenor de los resultados del análisis se han elaborado una serie de recomendaciones que deberían ser tenidas en cuenta por el Comité Directivo en las propuestas de actuación que recoja un futuro Plan de Impulso.

En el presente informe, estructurado según las áreas de interés mencionadas antes, se ha **recogido información para la elaboración de un análisis de Debilidades, Amenazas, Fortalezas y Oportunidades (a partir de ahora DAFO) de cada una de las áreas.** Los diferentes análisis se han recogido luego en un análisis DAFO del estado de las tecnologías del lenguaje en España.





5.

Recursos lingüísticos en
español y lenguas co-
oficiales

5. Recursos lingüísticos en español y lenguas co-oficiales



5.1. Estado del procesamiento lingüístico y traducción automática en español y lenguas co-oficiales

El procesamiento lingüístico incluye una amplia variedad de tareas que tienen como objetivo hacer explícita la información implícita que se deriva de la comprensión de un mensaje lingüístico. Estas tareas suelen desarrollarse por **herramientas de procesamiento lingüístico** de alta complejidad y especialización, que **enriquecen el texto con anotaciones** que hacen explícita la información que está implícita en los textos, como por ejemplo que la forma *compusieron* es la tercera persona del plural del pasado del verbo *componer*.

En un primer nivel tenemos tareas de pre-procesamiento o normalización de los textos de entrada, incluyendo **codificación de caracteres, estructura visual de los textos e identificación del idioma**. En un segundo nivel tenemos una serie de tareas de análisis a nivel de palabra, que incluyen la segmentación del texto en oraciones y palabras, y el análisis morfológico. En un tercer nivel se explicita la relación sintáctica entre las palabras de la oración, y se marcan las entidades nombradas y locuciones. En un cuarto nivel se realiza el análisis semántico, que incluye la desambiguación léxica respecto a una base de conocimiento externa, identificar la estructura argumental y los papeles temáticos (quién, qué, cómo) y representar formalmente el significado de las sentencias como proposiciones lógicas. Un quinto nivel, realiza el análisis discursivo, que incluye la resolución de la correferencia y anáfora, así como las relaciones entre las proposiciones y la estructura discursiva, incluyendo las relaciones temporales y opiniones.

Todas las tareas mencionadas anteriormente, incluso las de aparente sencillez, **son de difícil resolución y ninguna herramienta consigue resolver el 100% de los casos correctamente**. Son por tanto herramientas heurísticas. Aunque la aproximación dominante sea el hacer las tareas secuencialmente en un orden predeterminado, en casi todos los casos hay interdependencias que afectan a los

resultados. Por ejemplo, para segmentar oraciones es necesario decidir si el punto marca el final de la oración o es parte de un acrónimo.

La **aproximación dominante** para el desarrollo de estas herramientas consiste en desarrollar cada herramienta por separado usando el **aprendizaje automático supervisado**. Dada una tarea que debe producir cierta anotación lingüística, se proceden a realizar los siguientes pasos:

- 1) Diseñar y anotar grandes cantidades de texto manualmente con la información que debe aprender la herramienta. Por ejemplo, enriquecer las oraciones con árboles sintácticos explícitos, si la herramienta que se quiera construir es un analizador sintáctico que produzca, para cualquier sentencia, un árbol sintáctico. Esta es una tarea eminentemente lingüística. Si los textos no anotados son **corpus crudos**, los anotados manualmente son **corpus anotados**.
- 2) En algunos casos, es necesario desarrollar recursos lingüísticos adicionales, por ejemplo en forma de **bases de datos** léxicas que recogen todas las formas canónicas (o lemas) posibles de un idioma, o bases de conocimiento que contengan los conceptos relevantes del dominio. Esta es una tarea eminentemente lingüística.
- 3) Desarrollar **algoritmos** especializados de aprendizaje automático que aprendan a partir de los datos anotados manualmente y las bases de datos disponibles, para poder producir las anotaciones adecuadas para cualquier texto de entrada. Esta es una tarea eminentemente informática.
- 4) Dado que son aproximaciones heurísticas, es necesario **evaluar** las herramientas, y, en el caso de que el rendimiento no sea satisfactorio decidir si hacen falta más anotaciones (o un diseño distinto de la tarea de anotación) o algoritmos más sofisticados.

La aproximación alternativa es escribir reglas manuales (análogo al paso 1) y desarrollar algoritmos que ejecuten esas reglas (análogo al paso 3). En resumen, el proceso es muy **costoso**, ya que requiere un esfuerzo manual ingente (pasos 1 y 2), así como algoritmos avanzados de aprendizaje automático. Los **corpus crudos**, **corpus anotados** y **bases de datos conforman los recursos lingüísticos**, y los **algoritmos las herramientas de procesamiento lingüístico**.

Es de resaltar que el proceso ha de **repetirse para cada tarea, idioma y muchas veces dominio**, especialmente el etiquetado y desarrollo de bases de datos, pero también el desarrollo de algoritmos, dada las diferentes características de cada idioma.

La **calidad** de las herramientas de PLN para un idioma depende, por tanto, del tamaño y calidad de **los recursos lingüísticos** disponibles para ese idioma, y, en menor medida, de la mejor o peor adaptación de los algoritmos a las características de ese idioma.



Los estudios realizados por la Multilingual Technology Alliance (META) y la red de excelencia asociada META-NET, en el estudio Europe's Languages in the Digital Age señalan que la carencia de recursos para el procesamiento del lenguaje natural en una lengua particular son la barrera más importante para el desarrollo en Europa de aplicaciones de la industria del lenguaje.

Antes de proceder a un análisis del estado de las herramientas de PLN, es necesario hacer una lista de tareas de anotación estándar en PLN y TA:

1. Pre-procesamiento de textos de entrada

- Conversión entre juegos de caracteres
- Validación estructural de textos marcados
- Conversión entre formatos y estándares
- Limpieza de textos (por ejemplo, código html, código edición, etc.)
- Identificación de idioma
- Alineación de textos paralelos (traducción uno del otro) necesarios para aprender modelos de traducción automática

2. Análisis a nivel de palabra

- Segmentación de oraciones
- Tokenización, segmentación de unidades de análisis (palabras)
- Análisis morfológico (segmentar, reconocer la categoría gramatical y otros rasgos de las palabras)
- Lematización y desambiguación morfosintáctica (decidir una entre posibles categorías gramaticales y demás información morfológica, reconocer la raíz)

3. Análisis sintáctico

- Tratamiento de expresiones multipalabra, incluyendo locuciones y perífrasis
- Análisis sintáctico, ya sea como árbol de constituyentes o de dependencias
- Reconocimiento y clasificación de entidades con nombre

4. Análisis semántico

- Semántica léxica: Desambiguación de acepciones y entidades con respecto a bases de conocimiento
- Estructura argumental y papeles semánticos
- Representación del significado de la oración, como forma lógica por ejemplo

5. Análisis de discursivo

- Identificación de elementos temporales y sus relaciones
- Identificación de la factualidad y opiniones
- Resolución de anáfora y correferencia
- Estructura del discurso más allá de la oración



La lista pretende dar una idea de la complejidad de las tareas de análisis del PLN.

De los informes *Libro Blanco*⁹ de META-NET, en los que se evaluaba el estado de desarrollo de las principales tecnologías y recursos de PLN y TA para 31 lenguas europeas, se puede extraer una comparación con respecto a la lengua mejor dotada, el inglés, para el español. Para las lenguas co-oficiales, en distinto grado, el estado de desarrollo es aún menor. En la tabla siguiente se ofrece una visión panorámica de la situación de los recursos y herramientas disponible para el procesamiento del español. Se estima la **relación con respecto al inglés, que es en este momento el idioma que cuenta con mejor dotación** (fuente: META-NET, Libro blanco¹⁰). Los criterios referenciados se refieren a lo siguiente:

- **Cantidad:** recursos/herramientas catalogados y cómo se compara su cantidad con los disponibles para inglés. La cantidad muestra su uso en diferentes aplicaciones, relacionándose con la demanda.
- **Disponibilidad:** se puede disponer, reutilizar y explotar los recursos/herramientas con facilidad, por ejemplo son de código abierto, hay restricciones severas de utilización por la licencia, el precio u otros factores? La falta de disponibilidad muestra los nichos de mercado y las barreras al crecimiento.
- **Calidad:** han sido evaluados y cumplen con los criterios estándar de calidad, o un amplio uso en aplicaciones demuestra su calidad? Están mantenidos y actualizados? La calidad muestra la viabilidad de las aplicaciones que los usan.
- **Cobertura:** pueden ser utilizados para diferentes géneros (por ejemplo: prensa y textos médicos), variación lingüística, fenómenos lingüísticos (por ej.: diálogo y prosa), etc. O existen diferentes versiones cada una adaptada a una de las variedades anteriores? La cobertura muestra el coste de la inversión al cambiar la aplicación a otros casos.
- **Madurez:** pueden ser integrados en aplicaciones comerciales, sin necesidad de adaptaciones específicas al cliente? Son escalables? La madurez muestra el coste de la inversión para pasar de prototipo a aplicación comercial.
- **Sostenibilidad:** Diferenciaría, por ejemplo, actualizaciones dependientes de recursos humanos o de adaptación automática a nuevos inputs o herramientas/recursos que cumplen con estándares definidos vs. APIs o formatos propietarios, ad-hoc. La sostenibilidad muestra el coste del mantenimiento y el coste de ser integrado en una aplicación.
- **Adaptabilidad:** el coste de adaptar o extender las herramientas o recursos a nuevas tareas, aplicaciones o dominios. La adaptabilidad muestra el potencial de crecimiento.

Las estimaciones, en porcentaje con respecto al inglés, se han basado en catálogos (como META-SHARE y ELRA Universal Catalogue), información disponible gracias a asociaciones (SEPLN) e informes realizados por diferentes investigadores y expertos.



⁹ <http://www.meta-net.eu/whitepapers/e-book/spanish.pdf>
¹⁰ <http://www.meta-net.eu/whitepapers/e-book/english.pdf>

En la tabla, el primer tercio hace referencia a las herramientas y tareas que se han comentado antes. El segundo tercio hace referencia a las aplicaciones, que suelen requerir de herramientas maduras, y que luego se suelen adaptar como productos de usuario final. El tercer tercio hace referencia a los recursos que dan soporte a las herramientas.

	Cantidad	Disponibilidad	Calidad	Cobertura	Madurez	Integrable/ estandar	Adaptabilidad
Tecnologías del Lenguaje: Herramientas y Tecnologías							
Tokenizador, Análisis y generación morfológica y Desambiguación	50%	100%	90%	9%	90%	90%	70%
Análisis sintáctico,	50%	70%	70%	40%	40%	80%	70%
Sintaxis: reconocimiento de entidades	30%	90%	90%	80%	70%	90%	50%
Semántica léxica y oracional (Desambiguación de sentidos, estructura argumental y papeles semánticos)	20%	70%	50%	70%	70%	80%	70%
Semántica: desambiguación de entidades a bases de conocimiento	10%	90%	60%	40%	50%	90%	50%
Análisis discursivo: opinión/sentimientos	50%	10%	50%	50%	90%	90%	60%
Análisis discursiva (resolución de anáfora y coreferencia, inferencia, identificación de fechas/relaciones temporales)	30%	30%	70%	60%	20%	90%	50%
Procesamiento avanzado del discurso (más allá de la oración)	20%	20%	50%	50?	0%	10%	10%
Tecnologías del Lenguaje: Aplicaciones							
Clasificación automática de documentos	80%	80%	90%	90%	90%	90%	80%
Recuperación de información (indexación multimedia IR, crosslingual IR)	80%	80%	90%	80%	80%	80%	50%
Extracción de información. Minería de textos y extracción de eventos.	20%	20%	50%	20%	50%	60%	50%

Generación de texto	30%	90%	50%	80%	80%	90%	80%
Resumen, Respuesta a preguntas, tecnologías avanzadas de acceso a información	20%	50%	50%	10%	20%	20%	20%
Traducción Automática	80%	90%	70%	50%	80%	90%	70%
Infraestructura Lingüística (Recursos, Datos, Bases de conocimiento)							
Corpus de referencia	70%	50%	80%	80%	90%	50%	60%
Corpus sintácticos (treebanks, dependency banks)	25%	100%	80%	50%	90%	90%	20%
Corpus semánticos (Semcor, Propbank)	20%	100%	60%	20%	20%	80%	20%
Corpus de discurso	20%	100%	80%	20%	20%	90%	20%
Corpus paralelo y memorias de traducción	70%	60%	80%	40%	90%	90%	50%
Modelos de Lenguaje	90%	90%	90%	80%	90%	90%	80%
Léxico, diccionarios y terminología	80%	80%	80%	40%	90%	90%	20%
Gramáticas	50%	50%	80%	50%	50%	50%	50%
Thesaurios, WordNets, Propbanks	30%	80%	80%	20%	50%	90%	20%

Como ilustración de la mejor situación del Inglés cabe mencionar que las empresas que se dedican al procesamiento del Inglés tienen a su disposición por canales públicos de cadenas de procesadores lingüísticos (p.ej. CoreNLP¹¹ u OpenNLP¹²), recursos lingüísticos y semánticos (p.ej. WordNet¹³, FrameNet¹⁴ o VerbNet¹⁵), grandes corpus texto crudo (p.ej. ClueWeb o GigaWord), así como corpus de gran tamaño anotadas con información lingüística variada y de alta calidad (p.ej. Ontonotes). Esa **extensa colección de procesadores y recursos está disponible de forma libre o bajo pago de licencias asequibles**. Este



¹¹ <http://nlp.stanford.edu/software/corenlp.shtml>
¹² <https://opennlp.apache.org/>
¹³ <http://wordnet.princeton.edu/>
¹⁴ <http://framenet2.icsi.berkeley.edu/>
¹⁵ <http://verbs.colorado.edu/~mpalmer/projects/verbnet.html>

último punto sobre licencias ha sido clave para la disponibilidad de los recursos y facilitar el impacto en aplicaciones industriales.

El español tiene a su disposición de versiones análogas a las anteriores, tales como cadenas de procesadores lingüísticos (p.ej. Freeling¹⁶ o ixa-pipes¹⁷), recursos lingüísticos (p.ej. Spanish WordNet¹⁸, SenSem¹⁹, Adesse²⁰, Spanish FrameNet²¹ o Ancora²²), grandes masas de corpus (p.ej. Spanish GigaWord²³, Spanish Clueweb²⁴), así como corpus de gran tamaño anotados con información lingüística variada y de alta calidad (p.ej. CREA²⁵, Ancora²⁶). Todos ellos tienen **limitaciones con respecto al Inglés, tal y como se detalla en los porcentajes de la tabla**. Como ilustración podemos mencionar licencias restrictivas o no disponibilidad, tamaño reducido de los recursos, tamaño reducido de los corpus de anotación que repercuten en la calidad de las herramientas.

Respecto a la TA, la siguiente tabla (Fuente Libro Blanco META-NET) refleja la calidad de la traducción automática estadística entre diferentes pares de idiomas, que refleja la buena situación de las herramientas de este tipo para TA. Esta situación puede cambiar en poco tiempo, dado que las aproximaciones actuales tratan de incluir herramientas de análisis lingüísticos, y por tanto **aquellos pares de idiomas que dispongan de buenas herramientas dispondrían de mejor calidad de traducción**. Esto es especialmente relevante para el euskara, que al contrario de las lenguas romances, no dispone de idiomas tipológicamente cercanos, y además es un idioma de morfología aglutinante, fenómeno que dificulta y empeora los resultados de traducción (p. ej. ver finlandés en tabla).

	Idioma de destino – Target language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	-	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3	-	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	-	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	-	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	-	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	-	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	-	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	-	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	-	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	-	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	-	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	-	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	-	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	-	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	-	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	-	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	-	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	-	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	-	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	-	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	-	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	-

Hay que mencionar que la estimación de la calidad de estas herramientas es objeto **de campañas de evaluación que pone en marcha la comunidad**

- 16

17

18

19

20

21

22

23

24

25

26
- <http://nlp.lsi.upc.edu/freeling/>
<http://ixa2.si.ehu.es/ixa-pipes/>
<http://adimen.si.ehu.es/web/MCR>
<http://grial.uab.es/sensem/lexico?idioma=en>
<http://adesse.uvigo.es/>
<http://spanishfn.org/>
<http://clic.ub.edu/corpus/>
<https://catalog.ldc.upenn.edu/LDC2011T12>
<http://datahub.io/dataset/clueweb09>
<http://corpus.rae.es/creanet.html>
<http://clic.ub.edu/corpus/>



Respecto a las lenguas co-oficiales, mencionamos brevemente que la situación de las herramientas y recursos para el catalán es inferior al español. El euskara está en peor situación que el catalán y el gallego en peor situación que el resto.

internacional y en las que se establecen tareas, objetivos y materiales comunes para establecer comparaciones fiables entre herramientas. Estas evaluaciones permiten medir el progreso del área y tener indicadores de grupos, participantes nacionales y extranjeros. Estas campañas de evaluación han sido utilizadas en muchas áreas del PLN, empezando por las Message Understanding Conferences (MUC) financiadas por Defense Advanced Research Projects Agency (DARPA) para fomentar el desarrollo de mejores métodos de extracción de información, Document Understanding Conferences (DUC) y Text Analysis Conference (TAC) organizadas por el NIST de EUA. En Francia, también lo implantó el programa nacional Technolanguge-EVALDA e Italia lo está haciendo ahora EVALITA²⁷. También se han organizado tareas de PLN en el marco de SemEval²⁸ (en la última campaña se organizaron 16 tareas que incluían el inglés, 2 el español, 1 el árabe y otra el italiano), shared tasks de CONLL²⁹, CLEF³⁰, etc. También la SEPLN organiza periódicamente algunas tareas de evaluación. Por ejemplo, Tweet-Norm³¹ o TASS³².

5.1.1. Análisis DAFO

A continuación, se expone un **análisis DAFO** del procesamiento lingüístico y TA en español y lenguas co-oficiales:

Oportunidades

- Capacidad para desarrollar **recursos y herramientas para el PLN y TA** para el castellano y lenguas cooficiales al nivel del inglés, así como para el español de América.

Fortalezas

- Recursos y herramientas propias, **consolidadas y robustas** para hacer el procesamiento básico de PLN y TA para el castellano, catalán, vasco y gallego, además del inglés.

Debilidades

- Recursos y herramientas de PLN y TA para el castellano, catalán, vasco y gallego a un **nivel inferior al del inglés**.
- Inexistencia de recursos y herramientas para el **español panamericano**.
- **Escasa colaboración entre empresas** y entre empresas y grupos de investigación impide la reutilización de datos y herramientas y multiplica la



²⁷ <http://www.evalita.it/>
²⁸ <http://alt.qcri.org/semEval2015/>
²⁹ <http://www.cs.brandeis.edu/~clp/conll15st/>
³⁰ <http://clef2015.clef-initiative.eu/CLEF2015/>
³¹ <http://komunitatea.elhuyar.org/tweet-norm/>
³² <http://www.daedalus.es/TASS2015/tass2015.php>

inversión que realizan las empresas, restándole efectividad en otros ámbitos como la promoción comercial.

Amenazas

- **Desarrollo no español** de recursos y herramientas para el PLN y TA para el castellano y español panamericano.

5.2. Estándares, modelo de licencias, herramientas comunes

Con el término **recurso lingüístico** se hace referencia a conjuntos de datos lingüísticos (normalmente de gran tamaño) en formato electrónico anotados o enriquecidos para su uso en la construcción, mejora y evaluación de aplicaciones, sistemas, tecnologías y herramientas de procesamiento del habla y texto escrito. Son ejemplos típicos de recursos lingüísticos los corpus monolingües, paralelos o memorias de traducción, multilingües y multimodales, los léxicos o diccionarios, las gramáticas, las bases de datos terminológicas, las ontologías, etc.

La creación y uso de estos recursos son de gran importancia para un gran número de **aplicaciones avanzadas** tales como recuperación de información, traducción automática, etc.

Los **formatos en los que se presentan** estos recursos son muy variados. La clasificación realizada a continuación es útil para entender los derechos asociados con cada tipo de recurso. Se pueden citar los siguientes:

- **“Raw data”** (datos en bruto). Conjunto de datos recogidos tal y como han sido creados sin haber efectuado en ellos ninguna anotación.
- **Datos primarios.** Para algunos autores son el punto de partida para el procesamiento lingüístico; se distinguen de los anteriores en el hecho de que en ellos existe algún tipo de estructuración; no contienen anotación lingüístico pero sí pueden contener algún tipo de información estructural del documento.
- **Datos secundarios.** Aquellos que contienen anotaciones que añaden valor a los datos originales. Por ejemplo, los datos (documentos) se han podido estructurar en un corpus con sus correspondientes metadatos (origen, lengua, dominio) y pueden estar anotados a diferentes niveles desde el punto de vista lingüístico.
- **Recursos derivados.** Nuevos recursos derivados de otros ya existentes, por ejemplo como combinación de varios recursos existentes con datos primarios o secundarios: de corpus monolingües a corpus bilingües; o la inducción de un léxico a partir de un corpus de textos.
- **Corpus vs Base de datos.** Términos que, en general, se confunden. Normalmente hacen referencia a colecciones de datos secundarios usando “corpus” para referirse a colecciones de texto escrito y “base de datos” para referirse a grabaciones de voz, a colecciones de términos, etc.

Conviene recordar la distinción entre los términos recursos lingüísticos y herramientas, sistemas, aplicaciones y tecnologías lingüísticas. Comienza con el concepto de recursos lingüístico y más tarde se revisan los conceptos asociados al término tecnología lingüística.

Los metadatos, son parte del recurso, y se utilizan para la indexación y para facilitar la identificación del recurso. Incluyen información que describe el recurso que puede ser de varios tipos: i) administrativo (sobre creadores, contactos, etc.), ii) sobre el contenido (tipo de recurso, idioma, ...), información técnica (codificación, anotación, ...), iii) legal (licencia, ...). Algunos apartados de los metadatos pueden ser confidenciales.

Refiriéndonos a las “**herramientas**”, “**tecnologías**”, “**sistemas**” o “**aplicaciones lingüísticas**”, en primer lugar habría que indicar que es difícil distinguir entre estos términos. Algunos autores identifican una herramienta con un componente básico técnico (normalmente software) que produce una transformación sobre los datos o inserta información adicional (anotaciones). La combinación de herramientas básicas crea componentes tecnológicos que pueden integrarse con otros para dar lugar a nuevas tecnologías que a su vez se pueden incluir en aplicaciones.

La producción de un recurso lingüístico se realiza en **diferentes fases** que conviene conocer, ya que cada una presenta una problemática concreta con respecto a la elección de licencia. Así se distinguen las siguientes fases:

1. Tarea de especificación: implica la declaración y el intercambio de ideas y conocimiento común sobre know-how. Las especificaciones pueden necesitar algún tipo de licencia pero no parece muy probable ya que el copyright protege lo que se identifica como “*original works of authorship that are fixed in any tangible médium of expression*”.

2. Tarea de producción: compilación de datos primarios o inducción o explotación de nuevos datos a partir de datos primarios o secundarios. Se necesita el acuerdo de los creadores de los recursos originales. Las restricciones de uso y distribución definidas en los recursos originales marcarán las de los recursos derivados.

3. Tarea de validación: se refiere a procedimientos sofisticados de evaluación y control de calidad. En el caso de que los procedimientos de control sean realizados por personas diferentes a los creadores y dispongan de una copia del recurso, han de ajustarse a las obligaciones estipuladas en la cesión del recurso con el objetivo de la validación

4. Tarea de distribución: puede ser libre o no, requiere un acuerdo previo con los/las creadores/as del recurso.

5. Tarea de mantenimiento: tiene como objetivo las mejoras, ampliaciones, y la resolución de errores. Es importante obtener la autorización inicial de los/as creadores/as del recurso.

En resumen, cada tarea de las descritas tiene **asociados sus requerimientos con relación a responsabilidades** con los creadores del recurso.

Todas las fases descritas en este punto plantean diferentes cuestiones sobre la **propiedad intelectual y las maneras de licitar** los recursos y herramientas lingüísticas que se generan.

5.2.1. Modelos de licencias

Las licencias de los recursos lingüísticos han de incluir el **derecho a consultar el contenido de cada recurso lingüístico**, pero también los de **transformación**,



Es necesario distinguir entre licencias para la distribución de software y licencias para la distribución de contenidos digitales.

difusión y distribución de dichos recursos o de sus derivados, siempre de acuerdo con las necesidades y las políticas de los/las creadores/as y usuarios/as de los recursos.

Con respecto a las licencias para distribución de software, conviene señalar que hay **más de 2.000 tipos de licencias de software diferentes**. De hecho, cada desarrollador puede “crear” su propia licencia. En todo caso, estudios cuantitativos recientes han establecido que **alrededor del 90% de los proyectos de software abierto usan unos 10 tipos de licencias**³³. En general es posible distinguir dos tipos de licencias de código abierto:

- **Licencias permisivas:** tipología de licencias que, en general cuentan con pocas condiciones para el uso del software. Las tres licencias más populares según diversas fuentes son las licencias MIT, BSD y Apache License 2.0
- **Licencias recíprocas o copyleft:** tipo de licencia que establece como condición que cualquier software derivado de otro software debe ser distribuido con la misma licencia que el original. Las licencias más conocidas de este tipo son la GPL (versiones 2 y 3) y la LGPL. La característica más definitoria de la licencia GPL es la condición de que cualquier variante o derivado de un software bajo licencia GPL debe ser a su vez redistribuido con la misma licencia que el software original, es decir, GPL.

Como resumen cabría indicar que el reciente estudio de BlackDuck se muestra una **evolución del software abierto desde las licencias recíprocas hacia las licencias permisivas**³⁴. Aunque la licencia GPL sigue siendo la más usada según este estudio, cabe destacar que, entre los proyectos monitorizados en este estudio, las versiones 2 y 3 de GPL son usadas por alrededor del 37% de los proyectos, mientras que si se suman las tres licencias permisivas arriba mencionadas, representan el 42% de los proyectos de software.

Esta evolución puede deberse a una **variedad de factores**, entre los que se encontraría la mayor facilidad para los desarrolladores de usar las licencias permisivas, la influencia de comunidades forjadas alrededor del software libre y/o de código abierto como la fundación Apache, así como el tradicional rechazo del software comercial a las licencias recíprocas que pueden ser vistas como contaminantes ya que obligan a abrir el software en el que se integra un componente con licencia recíproca.

Así, las licencias permisivas, a la vez que garantizan la **disponibilidad pública del software distribuido**, permiten **usar el software original para el desarrollo o creación de software derivado** que pueda ser distribuido y/o comercializado,



En primer lugar, el software desarrollado con proyectos financiados con dinero público sigue siendo público para el beneficio de la sociedad en su conjunto y, en segundo lugar, las empresas y centros tecnológicos pueden realizar derivados de ese software y elegir el modelo de licencia que se adecue más a sus necesidades de innovación y competitividad en su sector tecnológico.

³³ https://www.upf.edu/innovacio/_pdf/UPF_2013_presentacix_M.Bain.pdf

³⁴ <http://redmonk.com/sogrady/2014/11/14/open-source-licenses/>

en su totalidad o en partes, bajo una licencia diferente, más acorde quizá al modelo de negocio tradicional. Esta característica de las licencias permisivas las convierte en candidatas ideales para la transferencia de tecnología de PLN de la universidad a la empresa.

La propiedad intelectual

La propiedad intelectual, según la definición de la Organización Mundial de la Propiedad Intelectual (OMPI), se refiere a **toda creación de la mente humana**. Los derechos de propiedad intelectual protegen los intereses de los creadores al ofrecerles prerrogativas en relación con sus creaciones. La propiedad intelectual tiene que ver con las creaciones de la mente: los inventos, las obras literarias y artísticas, los símbolos, los nombres, las imágenes, los dibujos y modelos utilizados en el comercio. La propiedad intelectual incluye dos categorías:

- La propiedad industrial que incluye **invenciones, patentes, marcas, dibujos y modelos industriales e indicaciones** geográficas de procedencia.
- El derecho de autor que abarca **las obras literarias y artísticas**, tales como las novelas, los poemas y las obras de teatro, las películas, las obras musicales, las obras de arte como los dibujos, pinturas, fotografías y esculturas, y los diseños arquitectónicos (en inglés, "Copyright").

Los derechos de autor incluyen los relacionados con las actuaciones de los artistas, productores de fonogramas y los concernientes a los productores de programas de radio y televisión. De esto se puede deducir que los derechos asociados a los recursos lingüísticos, entendidos como "trabajo original", son los correspondientes a derechos de autor. En Europa se inició la regulación del Copyright en 2011 y se dio un periodo de tiempo para que cada país adaptará su normativa a las directrices europeas. La mayor parte de los países europeos han adaptados su **legislación según las directrices europeas**. Es difícil aplicar a estos conceptos a los desarrollos propios del área de las tecnologías lingüísticas: recursos y herramientas.

Patentes

En el informe "META-NET: A network of Excellence forging the Multilingual Europe Technology Alliance" del 24 de febrero de 2012 se indica explícitamente que **algunas invenciones tecnológicas tales como máquinas de traducción automática, algoritmo de reconocimiento de habla, por ejemplo, podrían patentarse**. En el mismo informe se añade que la posibilidad de patente podría aplicarse también a los procesos de producción del recurso incluidos los procedimientos tecnológicos.

Directrices europeas sobre contenidos: bases de datos

En el campo de las tecnologías lingüísticas, el término *base de datos* se utiliza para denotar un **recurso lingüístico o una colección de recursos lingüísticos**



Las patentes se refieren a un derecho reconocido a:

"who invents or discovers any new and useful process, machine, article of manufacture or composition of matter, or any new and useful improvement thereof".

que no tienen que estar necesariamente organizados en bases de datos tal y como se las define en el mundo de la informática. Se refiere a una **compilación o recopilación estructurada de información**.

La Comunidad Europea introduce un nuevo tipo de derecho para los creadores de bases de datos, sean o no novedosas, diferenciándose de lo establecido para otros tipos de propiedad intelectual a los que se les pide básicamente que sean novedosos y originales. Un aspecto esencial que tiene que investigarse con cuidado es el hecho de que **los derechos del creador pueden ser “transferidos, asignados o cedidos”** bajo contrato, pero también que estos derechos tienen una vigencia de 15 años desde la fecha en que la base de datos se completó (el copyright tiene un tiempo de vigencia de 70 años).

Modelos de licencias en META-SHARE

La iniciativa europea Meta-Share, que consiste en la creación de un **catálogo de recursos lingüísticos en Europa**, con acceso directo a los mismos, ha adoptado el paradigma **Creative Commons** como base para el desarrollo de su oferta de licencias para la distribución de recursos lingüísticos.

Las licencias Creative Commons permiten establecer el **límite del uso y redistribución de los recursos lingüísticos originales**. Las licencias Creative Commons permitirían licenciar los recursos lingüísticos con una licencia tipo permisiva equivalente a las previstas para software, CC-BY o con una licencia de tipo “recíproca”. Las implicaciones de la decisión con respecto al proceso de transferencia de tecnología vendrían a ser las mismas que para las licencias de software permisivas y recíprocas.

El sistema de licencias modelo META-SHARE sigue la filosofía de la cultura del software abierto y se organiza en los **siguientes ejes**:

- Recursos lingüísticos que no cuentan con **ningún tipo de licencia** y permiten utilizar el recurso sin ninguna restricción y de manera permanente.
- **Creative Commons** (incluyendo todas sus variantes resultado de atribuciones diferentes).
- **Licencias META-SHARE Commons**: diseñadas para los miembros de la red META-SHARE que permiten hacer los recursos accesibles a otros miembros de la red. Se permite el uso y explotación de los recursos siendo el propietario del recurso el que controla su distribución. Este tipo de licencias también son adecuadas para el caso de distribución de los recursos fuera de la comunidad. Se ofrece un conjunto de plantillas de documentos legales (no licencias) diseñado para ayudar a todos los interesados (propietarios de los recursos, distribuidores y usuarios finales) a trabajar en un entorno amigable y transparente. Estas plantillas incluyen documentos tales como:

– Acuerdo del Creador del recurso (Depositor's Agreement, DA).



Para facilitar a los diseñadores y usuarios de los recursos lingüísticos la selección de licencia, META-SHARE describe y agrupa una serie de opciones que representan configuraciones de licencias recomendadas.

- Un Memorando de Entendimiento (MoU) para los miembros de la Red
- Un Acuerdo de Nivel de Servicio (Service Level Agreement, SLA). Una gran mayoría de las LRs disponibles actualmente desde los repositorios se ajustan a este modelo que fija las condiciones de uso para fines de investigación y comerciales.

5.2.2. Estándares en Tecnologías Lingüísticas

5.2.2.1. Uso de los estándares existentes

La introducción y utilización de estándares **produce grandes beneficios en la academia e industria** en la medida en que incrementa la utilización de datos, la combinación de herramientas y recursos lingüísticos en nuevos flujos de trabajo y, en general, el desarrollo de infraestructura.

Hoy en día existen **multitud de propuestas** para la anotación lingüística en diferentes niveles como son la **anotación textual de documentos** incluyendo segmentación básica en tokens y sentencias (Linguistic annotation framework LAF, Text Encoding Initiative, TEI³⁵), o la **anotación morfológica y morfosintáctica** (Morpho-syntactic Annotation Framework MAF³⁶; AWA³⁷), sintáctica (SynAf³⁸, TIGER³⁹) o semántica (SemAF, ISO-TimeML³).

Los estándares de anotación también cubren aspectos como la anotación de léxicos o diccionarios (Lexical Markup Framework, LMF⁴) y bases de datos terminológicas (TMF, Terminological Markup Language⁵), memorias de traducción (TMX), etc.

El organismo ISO ha reconocido varias de las propuestas de anotación existentes. Así, el organismo ha creado los estándares ISO/TC37, un conjunto de especificaciones para la anotación lingüística en varios niveles, incluyendo la segmentación de palabras (ISO 24614), anotación morfo-sintáctica (ISO 24611, ISO 24612) así como sintaxis (ISO 24615) y semántica (ISO 24617-1). Todos ellos están en proceso de revisión, lo que indica la inestabilidad debida a los cambios de la tecnología, fundamentalmente.



El estudio y definición de estándares para la anotación lingüística es un área activa de investigación desde que comenzara en la década de los 90 del siglo pasado. Desde entonces se han realizado grandes esfuerzos para establecer estándares que ayuden a la creación, explotación e integración de recursos de PLN.

³⁵ <http://www.tei-c.org/Guidelines/P5/>

³⁶ http://www.iso.org/iso/catalogue_detail.htm?csnumber=51934

³⁷ X. Artola, A. Díaz de Ilarraza, A. Soroa, and A. Sologaitoa. 2009. Dealing with complex linguistic annotations within a language processing framework. IEEE Transactions on Audio, Speech and Language Processing, 17(5):904–915.

³⁸ Declerck, Thierry. 2008 A framework for standardized syntactic annotation., Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008

³⁹ S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. 2013. Integrating NLP using linked data. In 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia.

Una descripción unificada y acordada para la representación de la información lingüística **simplifica enormemente la creación e integración de recursos de PLN**, permite que diferentes componentes tecnológicos sean interoperables reduciendo la creación de procesos de trabajo que implican varias herramientas. La introducción y utilización de estándares produce pues grandes beneficios en la academia e industria en la medida en que incrementa la utilización de datos, la combinación de herramientas y recursos lingüísticos en nuevos flujos de trabajo y, en general, el desarrollo de infraestructura, es decir, elementos de uso común.

Para facilitar la interoperabilidad, se han propuesto **esquemas que engloban múltiples niveles de anotación**. Los más conocidos son:

- **ATLAS**⁴⁰: que utiliza un formalismo común para la anotación lingüística llamado grafo de anotaciones (Annotation Graphs, AG). Los AG consideran las anotaciones siguiendo una estructura de grafo, donde las aristas dirigidas representan relaciones causales entre elementos.
- **PAULA** (Potsdam Interchange Format for Linguistic Annotation) ⁴¹, se trata de un formato XML multi-capa para representar múltiples niveles de anotación lingüística;
- **GATE** es una arquitectura y entorno de desarrollo para aplicaciones en el área de la ingeniería lingüística, basado en el formato de anotación de GATE está basado en TIPSTER.⁴²
- **UIMA**, un entorno de desarrollo de aplicaciones de PLN originariamente desarrollado por IBM hoy está disponible en Apache Software Foundation, tiene un sistema particular de anotación.

Más recientemente,⁴³ presenta NIF con el objetivo de utilizar principios y paradigmas **"Open Link Data" para anotar información lingüística**. NIF está basado en RDF y su principal objetivo es el de proporcionar un esquema uniforme de representación para los diferentes elementos que forman parte de la anotación lingüística. Así, NIF no especifica un esquema uniforme para representar análisis lingüísticos en diferentes niveles. Por ejemplo, NIF permite la representación de elementos ambiguos como anotaciones alternativas, pero no está claro cómo acceder a estas representaciones alternativas en los niveles superiores de análisis.

⁴⁰ S. Bird, D. Day, J. Garofolo, H. Henderson, C. Laprun, and M. Liberman. 2000. ATLAS: A flexible and extensible architecture for linguistic annotation. In Proc. of the Second International Conference on Language Resources and Evaluation, pages 1699–1706, Paris (France).

⁴¹ S. Dipper and M. Götze. 2005. Accessing heterogeneous linguistic data. generic XML-based representation and flexible visualization. In V. Saraswat and K. Ueda, editors, In Proceedings of the 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, pages 206–210. MIT Press, Poznan, Poland.

⁴² R. Grisham, Ed., Tipster Phase II Architecture Design Document. New-York University, 1995.

⁴³ E. König and L. Wolfgang. 2000. A description language for syntactically annotated corpora, Proceedings of the 18th conference on Computational linguistics - Volume 2, COLING '00, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1056–1060.



Una de las razones para desarrollar el formato NAF en el proyecto NewsReader fue, simplemente, la inexistencia de estándares de anotación para representar aspectos lingüísticos tales como la polaridad en sentencias y sintagmas o el marcado explícito de la atribución de expresiones textuales, etc

En los últimos tiempos se ha trabajado activamente en la **elaboración de los formatos KAF y NAF**, desarrollados en los proyectos europeos KYOTO y NewsReader, respectivamente. KAF/NAF está basado en varias propuestas ISO, y ofrece un esquema de representación XML de tipo "stand-off" para anotaciones lingüísticas en múltiples niveles, desde la tokenización hasta niveles semánticos de representación de entidades, predicados y eventos. KAF/NAF es usado satisfactoriamente para anotar documentos en varios idiomas (español, francés, holandés, italiano...) y actualmente existe una gran variedad de procesadores lingüísticos (IXA-Pipes, Freeling) que consumen y producen documentos en este formato.

5.2.2.2. Lenguas de estandarización

Si bien existen varias propuestas e iniciativas, lo cierto es que, a día de hoy, **no existe una única anotación lingüística** que sea ampliamente adoptada en el ámbito académico o industrial. Las razones de ello son numerosas, pero se podrían resumir en lo siguiente:

- Los objetivos de los estándares lingüísticos son muy variados, y muchas veces contradictorios. Así, **un estándar lingüístico ha de ser flexible para poder representar una gran variedad de fenómenos lingüísticos**; por otro lado, el estándar ha de ser **rígido** de tal manera que los desarrolladores de aplicaciones sepan de forma unívoca cómo se anota la información.

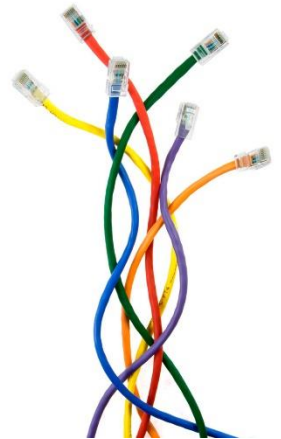
El PLN es un **área muy dinámica en la que, continuamente, aparecen nuevas técnicas y aplicaciones**. Así, es muy difícil, si no imposible, prever una representación unificada que contemple todos los niveles de análisis requeridos en las aplicaciones actuales de PLN.

- A pesar de que los estándares han de servir al PLN, algunos estándares de anotación han sido **desarrollados como ejercicio teórico de explotabilidad** pero carecen de herramientas, librerías y/o APIs para el desarrollo de aplicaciones que gestionen documentos anotados siguiendo ese formato.
- El hecho de que muchos estándares hayan sido diseñados a partir de datos del inglés **dificulta su adopción para otras lenguas** con características morfológicas y/o estructurales diferentes.

5.2.2.3. Herramientas comunes

En este apartado se van a identificar los diferentes niveles de trabajo a partir de una lista de las tareas comunes básicas necesarias para cualquier desarrollo en el ámbito de las tecnologías del lenguaje descrita en la sección 1.1: Estado del procesamiento lingüístico y TA en español y lenguas co-oficiales:

- **Análisis del texto:**



- **Identificación de idioma.**
- **Tokenización:** tokenizar o segmentar es el proceso de partir un texto en elementos denominado tokens (unidades indivisibles). Un tokenizador o segmentador es el primero de los componentes que se utiliza en el procesamiento de texto. Si bien hay muchos tokenizadores disponibles, siempre es necesario adaptarlos para objetivos o dominios específicos.
- **Nivel morfológico:**
 - **Segmentación morfológica** que permite separar la palabra en sus componentes morfológicos, la raíz (lexema) y los morfemas de flexión o derivación.
 - **Análisis morfológico**, proceso que da como resultado las posibles interpretaciones morfológicas de una palabra
 - **Lematización** que consiste en la asignación, en forma de etiqueta, de lema (o forma canónica) a una palabra tal y como se encuentra en el discurso textual. La lematización aporta una mayor flexibilidad en las búsquedas en textos.
 - **Etiquetado de categorías gramaticales** (conocido POS Tagger) que consiste en la asignación de la categoría gramatical a cada palabra de un texto considerando su posición en el texto real. Esta actividad suele llevar incluida la tarea desambiguación morfológica. En muchas ocasiones la lematización y el etiquetador de categorías gramaticales se realizan de manera conjunta.
 - **“Stemming”** proceso por el que se truncan las palabras de los documentos con el fin de obtener el lema; este proceso es muy común para el tratamiento de lenguas con poca flexión (inglés) pero no es adecuado para lenguas como el euskara;
- **Nivel sintáctico:**
 - **División de un texto en frases**, este proceso tiene que considerar problemas tales como los planteados por los signos de puntuación que no constituyen una frontera de frase.
 - **Identificación de sintagmas** (chunking) cuyo objetivo es detectar cadenas nominales, preposicionales y verbales.
 - **Análisis sintáctico** que intenta determinar la validez de una expresión (normalmente respecto a una gramática) y de obtener su estructura sintáctica. Este proceso da como resultado un árbol sintáctico, que representa la estructura de constituyentes de una frase o la estructura de dependencias de la misma.
 - **Identificación de entidades** nombradas.
 - **Clasificación de entidades** nombradas.
 - Reconocimiento de **expresiones multipalabra**;



La **European Language Resources Association** ("ELRA"- <http://www.elra.eu/>) es la promotora del BLARK y de ELDA, la distribuidora europea de un gran número de recursos lingüísticos y herramientas enmarcados en los diversos campos de la tecnología del lenguaje, gracias a la contribución de los distintos actores de esta comunidad. La misión de ELRA (como la de su imagen en EEUU Linguistic Data Consortium, www.ldc.upenn.edu) desde su fundación en 1995 ha sido **ayudar en la identificación y acceso a los recursos lingüísticos existentes** y facilitar su uso en diferentes ámbitos de las tecnologías de la información y comunicación. Esta asociación resulta eficaz para evitar que los investigadores y desarrolladores inviertan esfuerzos en volver a construir recursos ya existentes.

A través de la Asociación se han **distribuido recursos sobre lenguaje hablado, escrito, recursos terminológicos y multimedia**. Ha abordado y propuesto soluciones para problemas de logística comercial (precios, tasas, cánones), legal (licencias, derechos de propiedad intelectual) y de difusión. Los recursos que distribuye ELDA se encuentran ampliamente presentados en su catálogo de recursos (<http://catalog.elra.info/index.php>)

Afortunadamente, en los últimos años la comunidad que promociona la filosofía de software de código abierto se ha ido **ampliando de manera exponencial** y, en la actualidad, el **código abierto ofrece bibliotecas de programas y herramientas de gran calidad** que permiten resolver problemas típicos de procesamiento de textos. La existencia de estas herramientas facilita la innovación ya que permite aprovechar recursos.



En general, estas herramientas sirven para **resolver las tareas más comunes de PLN** tales como la tokenización, segmentación morfológica, segmentador de frases, análisis sintáctico, identificación y (en ocasiones clasificación) de entidades nombradas, la resolución de la correferencia. Algunos de estos paquetes, como por ejemplo OpenNLP (<http://opennlp.apache.org/index.html>), también incluyen algoritmos de máxima entropía y perceptrón basados en aprendizaje automático.

Algunos de estos paquetes ofrecen herramientas para distintos idiomas. Un ejemplo de ello son Stanford CoreNLP (<http://nlp.stanford.edu/software/corenlp.shtml>) y NLTK (<http://www.nltk.org/>), pero básicamente se han desarrollado **solo para el tratamiento del inglés, idioma para el que se dispone de grandes cantidades de corpus**, base indispensable para un buen funcionamiento de estas herramientas. Para el procesamiento de español y lenguas co-oficiales (aunque también incluye muchas otras lenguas europeas), el paquete de herramientas más ampliamente utilizado es Freeling (<http://nlp.lsi.upc.edu/freeling/>) con más de 100.000 descargas desde que se puso

En la página web <http://opensource.com/business/15/7/five-open-source-nlp-tools> se incluyen algunos recursos útiles para el desarrollo de aplicaciones más avanzadas en las que la comunicación o el acceso a la información sean un requisito fundamental.

a disposición como software libre en 2004. Por su parte, las ixa-pipes (<http://ixa2.si.ehu.es/ixa-pipes>) están disponible desde 2014.

5.2.3. Análisis DAFO

A continuación, se expone un **análisis DAFO** en relación a de los estándares, modelo de licencias, y herramientas comunes:

Fortalezas

- Incremento de las **directivas y ayudas** para el establecimiento del proceso de licencias (ejemplo Meta-Share)
- **Experiencia en iniciativas** tales como ELRA, META-SHARE, CLARIN-ERIC,
- **Experiencia en proyectos** europeos tales como OPENER, NEWSREADER, QT-Leap
- **Experiencia de grupos de investigación** en proyectos europeos en los que la definición de estándares ha sido un soporte metodológico básico y exitoso.
- Adecuada **representación de los grupos de investigación** en las comisiones de definición de estándares.
- Disponibilidad de **herramientas PLN multilingües** integradas mediante el uso de estándares.

Debilidades

- Todavía **no se ha llegado a un alto nivel de compartición** de herramientas de amplia cobertura entre las empresas y la academia.
- **Desconocimiento** preciso por parte de los desarrolladores sobre el **proceso de establecimiento de licencias**.
- **Desconocimiento** preciso por parte del **tejido empresarial** sobre el proceso de establecimiento de licencias.
- Dificultad para las empresas de tamaño pequeño y mediano aceptar un **modelo de licencias abierto** por el coste que supone la inversión
- Falta de trabajo para la **identificar casos de negocio pernicioso** de proveedores, vendedores LT e integradores de Tecnologías de la información.
- Dificultad para **interoperar** con herramientas externas.
- Dificultad para la **difusión y adopción por terceros** de los esquemas desarrollados.

5.2.4. Recomendaciones



En relación a la información expuesta en el presente apartado, se establecen las siguientes **recomendaciones**:

- Tender a la **creación de una infraestructura núcleo** de filosofía abierta apoyada por la administración y diseñada para dar cabida a la colaboración entre organizaciones que participan en distintos puntos de la cadena de valor.
- Facilitar el proceso de **diseño de licencias** de explotación.
- Diseñar, desde el punto de vista comercial, **estrategias de mercado para la explotación** de las tecnologías lingüísticas y facilitar el establecimiento de alianzas.
- Diseñar, por parte de los desarrolladores, **productos modulares e incrementales**.

5.3. Dominios especializados

El PLN y la TA **deben adecuar sus herramientas y recursos a los diferentes usos que se hacen de una lengua** y a la introducción de nuevas palabras y nuevos usos de palabras. Los dominios especializados son la fuente más importante de nuevas palabras o voces. En España, el desarrollo de recursos para los dominios especializados está cubierto por las organizaciones y organismos que trabajan en el ámbito de la terminología y son cruciales para áreas como la producción y traducción de textos especializados. En esta sección se describen las principales fuentes de información.

5.4. Asociaciones terminológicas

Existen **varias asociaciones** tanto regionales, como nacionales e internacionales en torno a la terminología:

- [AET](#): Asociación Europea de Terminología
- [AETER](#): Asociación Española de Terminología
- [IITF](#): Instituto Internacional de Terminología e Investigación
- [INFOTERM](#): Centro Internacional de Información Terminológica
- [REALITER](#): Red Panlatina de Terminología
- [RITERM](#): Red Iberoamericana de Terminología
- [SCATERM](#): Societat Catalana de Terminología

La asociación más importante en España, y cuya representatividad es notoria, es AETER creada en el año 1997 en Madrid por iniciativa de personas e instituciones interesadas en el estudio sobre la terminología y los lenguajes de especialidad, y en la elaboración y consulta de recursos terminológicos en español y en las demás

ALTER se trata de una asociación de ámbito nacional, sin ánimo de lucro, cuyos objetivos principales son desarrollar y promover la Terminología como disciplina, estimular la elaboración de recursos terminológicos necesarios y difundir los

lenguas de España. Entre las **actividades más representativas** de AETER se pueden encontrar las siguientes⁴⁴:

- Organiza regularmente **Jornadas de Terminología** con la finalidad de facilitar el intercambio científico y concienciar a las entidades generadoras y consumidoras de terminología de la importancia que ésta tiene en la comunicación y la divulgación en ámbitos de especialidad.
- Mantiene relaciones con organismos, asociaciones y organizaciones autonómicas, nacionales e internacionales con el objeto de **favorecer la coordinación del uso común** de las lenguas especializadas.
- Forma parte de la **Asociación Europea de Terminología (AET/EAFT)**, de la Red Iberoamericana de Terminología (RITerm) y de la Red Panlatina de Terminología (REALITER). Es miembro, en relación de reciprocidad de la Asociación catalana de Terminología (ACATERM).

Son **miembros de AETER** personas a título individual e instituciones muy diversas, entre las que se cuentan la Real Academia de Ciencias Exactas Físicas y Naturales y la Real Academia Española, Departamentos e Institutos de distintas universidades, la Asociación Española de Normalización y Certificación (AENOR), el Centro de Información y Documentación (CINDOC) del Consejo Superior de Investigaciones Científicas (CSIC), los Centros de terminología de otras lenguas de España, TERMCAT, UZEI y TERMIGAL, y diversas entidades privadas del ámbito de las industrias de la lengua y la edición.

En cuanto a la recopilación de las voces especializadas en español de diversos ámbitos de conocimiento, **las Reales Academias juegan el papel más importante en España** ya que son las instituciones que representan la autoridad sobre la correcta aplicación de las palabras tanto de uso como de especialidad. Centrando en el ámbito especializado los trabajos lexicográficos y terminológicos que actualmente se encuentran en estas instituciones son los siguientes:



5.4.1. Real Academia Española

La RAE dispone del *Corpus científico-técnico* (CCT)⁴⁵, creado alrededor del año 2001 y formado por un núcleo de **tres millones de palabras y complementado con otras publicaciones científicas** como el manual MERCK de medicina. También dispone de algunos repertorios terminológicos extraídos de diccionarios y glosarios:

- Diccionario Esencial de las Ciencias II de la RACFYN
- Diccionario médico etimológico de MONTEFIORE
- Glosario de Saludalia

En la elaboración de esta magna obra han participado más de 120 ingenieros y expertos en los distintos campos de la Ingeniería con el apoyo de un equipo lexicográfico dotado de los mejores medios tanto personales como materiales.

⁴⁴ <http://www.aeter.org/>

⁴⁵ http://cvc.cervantes.es/lengua/esletra/pdf/02/016_ballester.pdf

- Neoloteca del Centre de terminología TERMCAT
- Diccionario terminológico y de siglas del subsector de lubricantes de la Asociación Española de Lubricantes
- Glosario médico multilingüe de la DG III de la Unión Europea
- Glosario oftalmológico de Oftalmored

Existe también un **acuerdo⁴⁶ firmado entre la RAE y el Consejo General del Poder Judicial (CGPJ)** para la asistencia en el proyecto del *Diccionario de términos jurídicos*, que cuenta con un equipo de colaboradores formado por al menos cincuenta profesores universitarios entre los que hay especialistas en cada una de las ramas principales en que el Derecho se divide académicamente. El CGPJ, por su parte, podrá proponer la incorporación a los trabajos de magistrados, jueces y letrados que, por razón de sus conocimientos y especialización, contribuyan al desarrollo y ejecución del diccionario.

5.4.2. Real Academia de Ingeniería

La Real Academia de Ingeniería publicó en el 2004 el *Diccionario Español de Energía* con un total de **12.000 voces** y en marzo del 2014 ve la luz el *Diccionario Español de Ingeniería* (DEI en adelante) que recoge más de **50.000 términos de la ingeniería** y que ha sido elaborado a partir de un corpus de referencia que contiene más de 1500 diccionarios y un corpus textual del que se han extraídos 112 millones de palabras de las distintas áreas de la ingeniería⁴⁷.

El *Diccionario Español de Ingeniería* recopila las **voces de la Ingeniería y de sus nueve grandes ramas**:

- Astronáutica, naval y transportes
- Agroforestal
- Construcción
- Tecnologías de la información y las comunicaciones
- Seguridad y defensa
- Química industrial
- Energía
- Ingeniería biomédica
- Ingeniería general

Estas nueve grandes ramas de la Ingeniería se han dividido en **treinta y dos campos** para recoger con precisión todas las voces técnicas utilizadas por los ingenieros afines a cada una de ellos.

⁴⁶ <http://www.rae.es/noticias/el-tribunal-supremo-y-la-rae-elaboraran-un-diccionario-de-terminos-juridicos>

⁴⁷ <http://diccionario.raing.es/>



Como última actividad, la Real Academia de Ingeniería ha comenzado a trabajar con instituciones y universidades de EE.UU e Iberoamérica para elaborar el Diccionario Panhispánico de Ingeniería.

Además de a la comunidad científico-técnica en general, se dirigirá, específicamente, a **tres grandes grupos**:

- **Especialistas**, es decir, ingenieros, técnicos, redactores técnicos y profesionales relacionados con la Ingeniería en general.
- **Especialistas en vías de formación**, es decir, estudiantes de grado y posgrado.
- **Mediadores lingüísticos**, es decir, traductores, intérpretes, periodistas especializados, expertos en comunicación, etc.

Destacar que es un **repertorio digital y abierto** a cualquier tipo de consulta. Además, Como novedad respecto a otros diccionarios electrónicos, la versión electrónica del Diccionario *Español de Ingeniería* es accesible al mayor número posible de personas, incluyendo a quienes tienen una discapacidad, a los mayores o a quienes acceden a ella haciendo uso de tecnologías poco convencionales.

5.4.3. Real Academia Nacional de Medicina

La Real Academia Nacional de Medicina publica en 2011 su **Diccionario de términos médicos (DTM en adelante)**. Este repertorio contiene **casi 52.000 entradas** que equivalen a unas 66.000 acepciones. Facilita 40.000 remisiones internas⁴⁸.

El DTM es un diccionario sobre el **léxico médico actual** (a pesar de que registra algunos términos históricos pero con vigencia en la actualidad o con nuevas acepciones) **del español peninsular** —si bien se anuncia en el prólogo que en futuras ediciones se recogerán las variantes del español de América de manera sistemática. El DTM incluye términos que se usan en textos médicos de diversa estructura, tanto monoléxicos como, sobre todo, poliléxicos; y todos reciben el mismo tratamiento terminológico, pues tienen una entrada autónoma.

El Diccionario de términos médicos incorpora equivalentes en inglés norteamericano y proporciona información etimológica e histórica para cerca de 7.000 tecnicismos médicos.

En resumen, las **características más importantes** del DTM son:

- Diccionario especializado
- De tamaño medio
- Edición impresa y en línea
- Enfoque terminológico
- Dirigido a médicos y a otros profesionales biosanitarios
- Importancia de la etimología
- Aportar el equivalente en inglés norteamericano de todos los términos definidos en el diccionario
- Carácter normativo y didáctico

⁴⁸ <http://dtme.ranm.es/index.aspx>



5.4.4. Real Academia de Ciencias Exactas, Físicas y Naturales

En 1983 aparece la primera edición del **Vocabulario científico y técnico** que recogía unos 13.000 vocablos. Siete años después, la RACEFyN, bajo la presidencia de Martín Municio, publica en 1990 con Espasa-Calpe la segunda edición del Vocabulario científico y técnico, con 751 páginas, que revisa y completa hasta unos 24.000 términos (unas 25.000 acepciones) el inventario del VCT1, y amplía el número de áreas consideradas. En 1996, la RACEFyN, bajo la misma presidencia, procede a publicar, esta vez con Espasa, la tercera edición del Vocabulario científico y técnico con 33.467 términos –unas 35 000 acepciones– e incorpora equivalencias español-inglés e inglés-español de todos los términos incluidos en el VCT, presentadas en listados independientes al final del mismo⁴⁹, inclusión de nuevas áreas (Biomedicina, Nuevos Materiales, Ingenierías), y mejor sistematización en otras (Botánica). En cuanto a las voces inglesas, su presencia está justificada por la demanda de muchos científicos, pero Martín Municio se apresuraba a afirmar que “ello no debiera impedir que en una lengua universal como la española estén presentes las lexías complejas, las sinonimias y los neologismos propios del lenguaje de la ciencia”.

El contenido del VCT elaborado podrá ser corregido, o modificado para su actualización continua, por los editores autorizados. El software adoptado permite también **la utilización de múltiples codificaciones de texto para la presentación de términos**, incluidos html, latex o texto plano junto a imágenes, animaciones y vídeos. Entre los cambios, mejoras e intenciones para esta versión digital, figuran:

- Eliminación de términos obsoletos.
- Actualización de las definiciones obsoletas.
- Introducción de términos nuevos de uso común en la ciencia actual.
- Equilibrio de las distintas áreas.
- Proceder con sumo cuidado en la adaptación de extranjerismos (españolizando cuando sea posible, y si hay varias opciones, eligiendo la más usada).
- Revisar, ampliar y actualizar las pautas lexicográficas de tratamiento del contenido científico y diseño de la estructura normalizada de las definiciones.
- Recuperar el protocolo de paso de términos por el filtro de la Comisión de Terminología Científica de la Academia.

5.4.5. Real Academia Nacional de Farmacia

En 2007 la Real Academia Nacional de Farmacia publica su **Diccionario Terminológico de las Ciencias Farmacéuticas** en su versión bilingüe

⁴⁹ http://www.rac.es/5/5_3.php

En la actualidad, el Vocabulario científico y técnico está publicado en medio digital mediante el software mediawiki –que procede de la wikipedia– es un software de código abierto que permite una consulta fácil y rápida de todos los términos del Vocabulario así como un abaratamiento de los costes respecto a otros softwares de carácter privado.



El Diccionario Terminológico de las Ciencias Farmacéuticas está publicado únicamente en papel por la editorial Ariel.

español/inglés. Este repertorio especializado consta de dos partes: la primera (inglés-español) tiene en torno a 15.000 entradas, y la segunda (español-inglés), más de 13.000.⁵⁰ Aunque en teoría el número de entradas de las dos partes debería ser igual, esto no es posible por las siguientes razones:

- No existe una relación biunívoca entre las dos lenguas.
- Lo que en una es relevante en la otra puede no serlo por distintos motivos.
- Lo que en inglés o en español es una unidad léxica, en la otra puede ser una expresión lingüística que desborde los límites de los que razonablemente se considera unidad léxica.

⁵⁰ <http://www.ranf.com/sesiones-y-actos/presentaciones/929-presentaci%C3%B3n-del-diccionario-terminol%C3%B3gico-de-las-ciencias-farmac%C3%A9uticas.html>

5.4.6. Bancos de datos terminológicos y glosarios

Cabe destacar la **importancia de los bancos de datos terminológicos y glosarios** ya que –en palabras de María Teresa Cabré– “[...] se constituían para extraer de ellos la nomenclatura de glosarios temáticos y diccionarios, o también para que fueran fuente de resolución de dudas lingüísticas sobre unidades terminológicas o léxicas en general tanto en contextos de traducción como en contextos de normalización lingüística. Los datos contenidos en estos recursos tenían el formato de registros de base de datos en los que a cada entrada léxica o terminológica se le asociaban tipos de información estructurada en campos. Los campos más representativos en los bancos terminológicos, por ejemplo, eran la categoría gramatical, el ámbito temático, la definición, y en el caso de bancos plurilingües, los equivalentes en otras lenguas.”

La Dirección General de Traducción de la Comisión Europea, en su página web, **recomienda los siguientes recursos** sobre dominios de especialidad:

- Agricultura, silvicultura, pesca y alimentación (AGROVOC)
- CILF: banco de datos terminológicos del Consejo Internacional de la Lengua Francesa (francés). Disponible en inglés, alemán y español
- Glosario de medio ambiente (AEMA)
- Electropedia (CEI): vocabulario electrotécnico en línea
- ESIS (Sistema Europeo de Información de Sustancias Químicas)
- ESLEE (Estudio de los Lenguajes Especializados en Español): vocabulario técnico y científico
- Glosario de política europea de competencia (947 kB) (también disponible en francés e inglés)
- EURYDICE: tesoro europeo de educación
- Euskalterm: Banco Terminológico Público Vasco
- FishBase: base de datos mundiales sobre peces
- Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO): bases de datos terminológicos, glosarios y tesauros
- Glosarios por idiomas
- IATE
- ILOLEX: base de datos de la OIT sobre las normas internacionales del trabajo
- ILOTERM: base de datos de la Organización Internacional del Trabajo
- Tesauro de la OIT
- Fondo Monetario Internacional (FMI): directorio multilingüe



Todos estos recursos son de acceso libre y constituyen una fuente de datos multilingües especializados de gran importancia para el PLN y la TA.



- Glosario de desarrollo regional: terminología de la UE
- Estadísticas: glosario multilingüe del Instituto Internacional de Estadística (IIE)
- Sitios de terminología de los socios de JIAMCATT (Reunión Anual Internacional sobre Traducción y Terminología Asistidas por Ordenador)
- TERMCAT (centro de terminología de la lengua catalana): diccionarios y bases de datos terminológicos en línea
- TERMINESP (banco de datos de la Asociación Española de Terminología)
- Terminometro (Unión Latina): terminología
- TERMITE (Unión Internacional de Telecomunicaciones)
- TERMIUM: Banco de datos terminológicos del Gobierno de Canadá
- TREMÉDICA: recursos de terminología médica y revista de traducción (Panace@)
- UNESCOTERM: base de datos terminológicos y de referencia
- UNTERM: base de datos terminológica multilingüe de las Naciones Unidas
- Diccionario multilingüe sobre el agua (AquaLingua)
- Base de datos de todas las aves del mundo
- WTOTERM: base de datos terminológicos de la Organización Mundial de Comercio
- MEDES (Buscador de información médica en español)

5.4.7. Otros recursos lingüísticos de interés para el PLN y la TA

A continuación, se exponen **otros recursos lingüísticos** de interés para el Procesamiento del Lenguaje Natural y la Traducción Automática no mencionados con anterioridad.

- La **Asociación de Traductores de la Administración Pública** (RITAP) mantiene la Glosateca, que compila los más importantes proyectos terminológicos que incluyen el español entre sus lenguas, desde las grandes bases de datos terminológicas institucionales hasta los numerosos trabajos elaborados por especialistas de organismos internacionales, universidades, centros de investigación u otras entidades, aplicando siempre un criterio de calidad. En el Libro Blanco de la Traducción e Interpretación Institucional en España, además de describir el marco jurídico de la traducción e interpretación institucional y un resumen de la situación de los traductores e intérpretes, entre otros temas, contiene un anexo con recursos documentales y terminológicos para la traducción institucional.

El corpus Iberia, que tiene unos 85 millones de palabras de textos obtenidos de las revistas de CSIC y de su repositorio institucional, se va ampliar su dominio médico con la incorporación 140 millones de palabra de artículos de revistas médicas electrónicas publicadas por Elsevier.

- El antiguo **Instituto de Estudios Documentales de Ciencia y Tecnología** (EDCYT) del CSIC, y anteriormente CINDOC, dispone también de algunos (pequeños) glosarios accesibles libremente. El CSIC además, tiene representación en el Comité Técnico de Normalización AEN/CTN 191 de Terminología. El grupo El español lengua de la ciencia (ELCI) centra sus objetivos generales en el estudio del español como lengua de transmisión del conocimiento especializado. Los integrantes de esta línea de investigación se ocupan fundamentalmente del análisis y descripción del léxico científico-técnico del español con fines lexicográficos. Para lograr sus objetivos ELCI cuenta con una Estación de Trabajo Lexicográfico (ETL) en la que se alojan herramientas como el corpus Iberia (el español, lengua de la ciencia en la sociedad del conocimiento), las bases de datos «Torres Quevedo» (diccionarios terminológicos del español) y Nebrija-Valdés –metalexicografía del español–, un tesoro terminológico a partir de una abundante producción terminográfica y, en fase de desarrollo, el observatorio de neología o del neologismo científico.
- **TERMINESP** es un banco de datos terminológico de la lengua española promovido por la Asociación Española de Terminología (AETER), creado a partir de los datos proporcionados por la **Asociación Española de Normalización y Certificación** (AENOR) y desarrollado en colaboración con la Fundación del Español Urgente (Fundéu BBVA), que lo incluye como recurso en la Wikilengua. El objetivo inicial de Terminesp es proporcionar las definiciones dadas en las normas españolas UNE de varias decenas de miles de términos, con equivalencias en otras lenguas, cuando las hay, tal como aparecen en ellas, sin más. Para ofrecer explicaciones adicionales, junto al término en español se da un enlace que permite crear una página en la Wikilengua con ese título y añadir la información que se crea oportuna, como explicaciones, precisiones, ejemplos, etc. TERMINESP está integrado en la Terminoteca de Wikilengua, desde donde se puede acceder a otros recursos.
- El **TERMCAT** es un centro de terminología en lengua catalana creado en 1985 por la Generalitat de Cataluña y el Institut d'Estudis Catalans. El TERMCAT tiene como misión el **desarrollo y la integración de la terminología catalana** en sectores especializado y en la sociedad en general. Mantiene un diálogo permanente con especialistas y usuarios y proporciona herramientas y recursos de calidad. Su actividad está enmarcada en el proceso de normalización de la lengua catalana y en un contexto global determinado por la sociedad del conocimiento, la diversidad y el multilingüismo.
- El **Centro Vasco de Terminología y Lexicografía (UZEI)** está conformado por un equipo que ofrece servicios de valor añadido en el ámbito de la lengua. Con una experiencia de más de 35 años, son un agente activo tanto en el ámbito del léxico especializado como en el del léxico general. El estudio y la gestión del léxico en euskera ha permitido ofrecer servicios lingüísticos de calidad. Hoy, además del euskera,

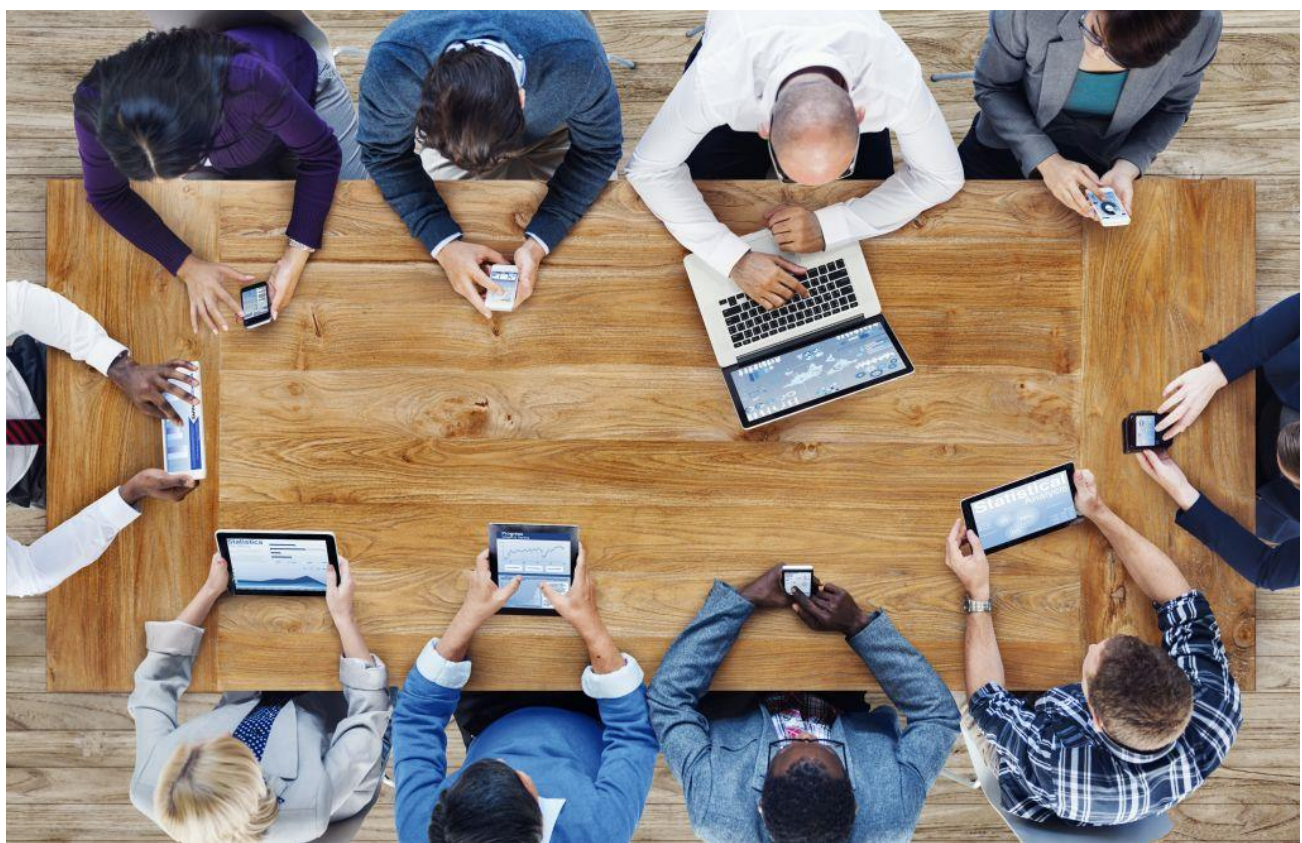


Día a día, el multilingüismo adquiere mayor importancia. Cada vez es mayor la demanda de creación, recepción e intercambio de información en varios idiomas, y las tecnologías de la lengua juegan un papel fundamental a la hora de atender esa demanda.

también gestiona y traduce otras lenguas. UZEI es una referencia indiscutible en el ámbito de la investigación y desarrollo de las tecnologías de la lengua, y ha creado numerosos recursos y herramientas en ese campo. Por ello, UZEI ha sido designado **miembro de la Red Vasca de Ciencia, Tecnología e Innovación**. También se han desarrollado otros proyectos en este ámbito, entre los cuales destaca euLEX, lexicón que constituye la base de todas las herramientas de PLN desarrollados por UZEI. En el ámbito del lenguaje especializado, UZEI ha sido un agente pionero en el trabajo terminológico desarrollado en el País Vasco. Creó el Banco de Terminología EUSKALTERM, y, desde que dicho banco de datos se convirtiera en banco público, ha sido responsable de su ampliación y mantenimiento. Asimismo, UZEI ha participado en varios proyectos internacionales en el ámbito de la terminología, desarrollados en colaboración con otros centros de terminología.

5.4.8. Conclusiones

Tras hacer el análisis previo de los dominios más representativos de las diversas Reales Academias y Asociaciones de traductores y terminólogos, se concluye que, para realizar una compilación normalizada y con autoridad que represente fielmente el idioma español en todos los ámbitos de conocimiento especializado, es necesario **crear una plataforma tecnológica del español que reúna a los mejores profesionales en lexicografía, terminología, lingüística aplicada e ingeniería lingüística** que elaboren un repertorio del español científico-técnico homogéneo y que dé impulso a la difusión de textos especializados en español.





6.

Situación de la
Investigación del
procesamiento lingüístico
en España e Iberoamérica

6. Situación de la Investigación del procesamiento lingüístico en España e Iberoamérica

La comunidad investigadora internacional en procesamiento del lenguaje natural (PLN) y traducción automática (TA) **compite a nivel mundial en el desarrollo de productos para el español como idioma internacional**. En EE.UU. y otras partes de Europa se han desarrollado herramientas de primera generación, por ejemplo de TA, que podrían aprovechar la oportunidad de mercado que representa una población de cuatrocientos o quinientos millones de hablantes. Además, en España, **un 30% de los ciudadanos tienen como lengua materna una lengua regional distinta del castellano**⁵¹, por lo que la investigación llevada a cabo a nivel nacional tiene en cuenta esa diversidad. La descentralización política ha permitido a las distintas regiones españolas fomentar la investigación y el desarrollo de la tecnología lingüística, como es el caso del catalán, el gallego y el vasco. El enfoque lingüístico intrínsecamente amplio de la investigación española, lejos de diluir el foco en las tecnologías del español representa un modelo a seguir para las políticas multilingües europeas. Por ese motivo, puede decirse que España es uno de los estados miembros con mayor potencial en tecnologías lingüísticas y en especial en TA.



6.1. Censo de instituciones y grupos de investigación en procesamiento de PLN y TA

6.1.1. Grupos de investigación

El primer censo de grupos de investigación en PLN en España fue confeccionado en 1997 por el hoy ya desaparecido Observatorio Español de Industrias de la Lengua (OEIL) del Instituto Cervantes⁵². En el censo de 1997 se inscribieron **21 grupos de procesamiento de lenguaje y voz vinculados a universidades, centros de investigación y a alguna división de alguna empresa privada**. Salvo

La investigación en PLN se lleva a cabo principalmente por informáticos, matemáticos, lingüistas y documentalistas, entre otros, muchas veces de manera inter y multidisciplinar.

⁵¹ http://eldiae.es/wp-content/uploads/2015/06/espanol_lengua-viva_20151.pdf

⁵² http://cvc.cervantes.es/lengua/anuario/anuario_98/listerri/p08.htm

las divisiones de IBM y Telefónica I+D, el resto de grupos censados por el OEIL continúan existiendo en la actualidad, con algún cambio de denominación en algún caso. La permanencia de la mayoría de estos grupos 18 años después debería interpretarse como un indicador positivo de la consolidación de estos grupos investigación en PLN en España. Muchos de los grupos de investigación españoles en PLN participan en redes temáticas como la Red TIMM⁵³ (La Red Temática en Tratamiento de Información Multilingüe y Multimodal), que es una red de investigación financiada por el Ministerio de Economía y Competitividad (TIN2011-13070-E). A continuación, se exponen los **principales grupos investigadores del territorio español**:

Galicia:

- Universidad de Vigo:
 - Grupo de investigación en Tecnoloxías e Aplicacións da Lingua Galega (TALG)
- Universidad de A Coruña:
 - Lengua Y Sociedad de la Información (LYS)

País Vasco:

- Universidad del País Vasco:
 - Grupo de Procesamiento de Lenguaje Natural (IXA)
- Universidad de Deusto:
 - LinguaMedia

Cataluña:

- Universidad Politécnica de Cataluña:
 - Grup de Recerca en Processament de Llenguatge Natural (GRPLN – TALP)
- Universidad Pompeu Fabra:
 - Tractament Automàtic del Llenguatge Natural (TALN)
 - Tecnologies dels Recursos Lingüístics (TRL – IULATERM)
 - Grup de Lingüística Computacional (GLICOM)
- Universidad de Barcelona:
 - Centre de Llenguatge i Computació (CLiC)
 - Grup de Recerca Interuniversitari en Aplicacions Lingüístiques (GRIAL)
- Universitat Oberta de Catalunya:
 - Language Processing Group (LPG)

De los datos obtenidos de la Red TIMM, y otras asociaciones como la SEPLN, en 2014 existirían más de 30 grupos de investigación en España, repartidos entre las diferentes universidades que trabajan en el reconocimiento del habla, el procesamiento del lenguaje natural, la traducción de texto a texto y la síntesis de voz. El censo de la Red TIMM de miembros individuales sobrepasa los 160 investigadores.

⁵³ <http://timm.ujaen.es/>



- Universidad Autónoma de Barcelona:
 - Laboratorio de Lingüística Informática (LaLI)

Comunidad Valenciana

- Universidad Politécnica de Valencia:
 - Grupo de Ingeniería del Lenguaje Natural y Reconocimiento de Formas (ELIRF)
 - Pattern Recognition and Human Language Technology (PRHLT)
 - Natural Language Engineering Lab (NLEL)
- Universidad de Alicante:
 - Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información (GPLSI)
 - Grupo Transducens

Andalucía:

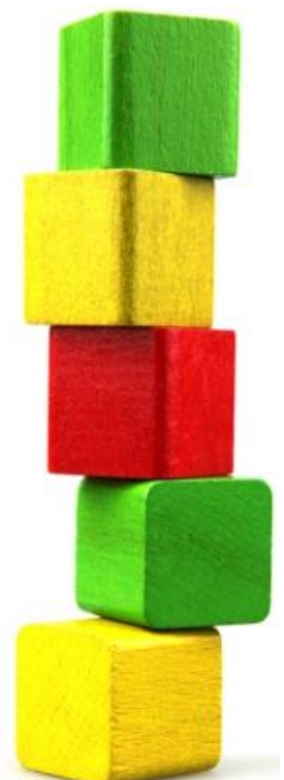
- Universidad de Jaén:
 - Sistemas Inteligentes de Acceso a la Información (SINAI)
- Universidad de Sevilla:
 - Grupo de investigación Julietta
 - Integración de Técnicas Automáticas en Lingüística y Compilación (ITALICA)

Asturias:

- Universidad de Oviedo:
 - Seminariu de Filoloxía Asturiana de la Universidá d'Uviéu

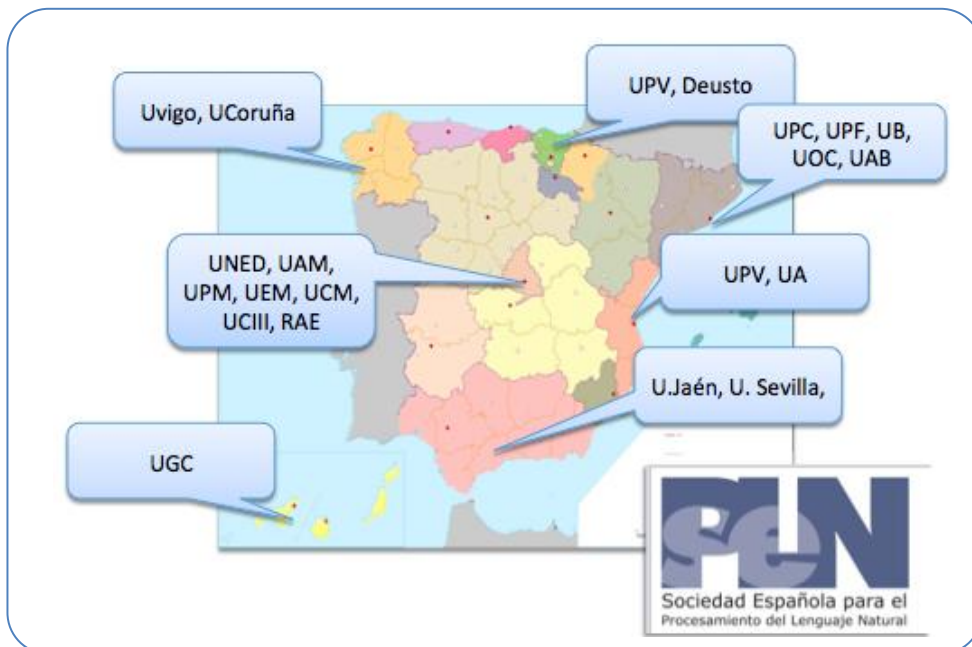
Madrid:

- Universidad Nacional a Distancia:
 - Grupo de Investigación en Procesamiento de Lenguaje Natural y Recuperación de la Información (NLP&IR)
- Universidad Autónoma de Madrid:
 - Laboratorio de Lingüística Informática (LLI)
- Universidad Europea de Madrid:
 - Departamento de Informática, Automática y Comunicaciones
- Universidad Complutense de Madrid:
 - Natural Interaction based on Language (NIL)



- Universidad Carlos III:
 - Laboratorio de Bases de Datos Avanzadas (LaBDA)
- Universidad Politécnica de Madrid
 - Grupo de Sistemas Inteligentes (GSI)
- Real Academia Española:
 - Centro de Estudios de la RAE

A continuación, se expone un gráfico en el que se muestra la **distribución geográfica de universidades e instituciones con los grupos de investigación más relevantes** en PLN en España, cuyo censo por comunidades autónomas sería la siguiente:



Según los informes META-NET **sobre el vasco, el catalán y el gallego hay para cada una de estas lenguas una comunidad de investigación en tecnologías del lenguaje** viable que ha producido recursos y tecnologías pero que están todavía lejos de las existentes para el español y mucho más para de las disponibles el inglés.

En las tres universidades del **País Vasco** hay **grupos de investigación en procesamiento y síntesis del habla, PLN y TA** que trabajan desde el año 1988 de manera coordinada. Actualmente participan en Langune, una asociación empresarial del sector de las Industrias de la Lengua de Euskal Herria, nacida en 2010 y que agrupa a más de 3 entidades de los campos de la traducción, contenidos, enseñanza y tecnologías del lenguaje.

En **Cataluña** destaca el **Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla (TALP)** de la Universidad Politécnica de Cataluña, que está



Las comunidades de investigación para el vasco, catalán y gallego no son suficientes para desarrollar el tipo de tecnología requerida para un verdadero soporte al multilingüismo.

especializado en PLN, tanto oral como escrito. Este centro se constituyó en 1998, está formado por un equipo de 50 personas y es miembro del Clúster Catalán de las Industrias de la Lengua (CLUSTERLINGUA), una asociación de compañías en Cataluña que proveen servicios relacionados con la multiculturalidad y el multilingüismo. TALP desarrolla y mantiene la herramienta de código abierto Freeling⁵⁴, que cuenta con más de 100.000 descargas y que procesa 10 lenguas, el español y las co-oficiales entre ellas. En el pasado existieron centros como el CREL (*Centre de Referència en Enginyeria Lingüística*, 1996-2000), gestionado por el *Institut d'Estudis Catalans* y la participación de las principales Universidades Catalanas, con el objetivo específico de promover la creación de herramientas y recursos para el procesamiento automático de textos en catalán en diversas aplicaciones.

El **Consorcio MAVIR** es una red de investigación co-financiada por la Comunidad de Madrid y el Fondo Social Europeo bajo los programas de I+D en TIC MA2VICMR (2010-2013) y MAVIR (2006-2009). El núcleo del consorcio MAVIR está formado por siete grupos de investigación de universidades y centros de investigación.

En el caso de la **lengua gallega**, los principales agentes implicados en la producción de recursos de tecnología lingüística son **los centros de investigación pública, fundamentalmente universitarios**. Entre los primeros, los grupos más activos y con más experiencia en esta disciplina son el Seminario de Lingüística Informática de la Universidad de Vigo integrado en el *Grupo de TecnoloXías e Aplicacións da Lingua Galega* (TALG) de esta universidad y el *Instituto da Lingua Galega* (ILG) de la Universidad de Santiago de Compostela; y, ya fuera del ámbito universitario, el Centro Ramón Piñeiro para la Investigación en Humanidades (CRPIH), que depende de la Xunta de Galicia.

Además, se ha identificado actividad en PLN en varios **países latinoamericanos con diferente grado de actividad** dependiendo del país. La Sociedad Norteamericana de Lingüística Computacional (NAACL) creó un pequeño fondo para regiones emergentes para establecer y cohesionar una comunidad de investigación sobre lingüística computacional que fomente las interacciones regionales entre estudiantes e investigadores. La asociación abrió un censo durante el período 2001-2013 donde constan más de 300 investigadores, algunos de ellos con Europa. La **distribución de las adscripciones por países** en el continente americano se detalla en la tabla 4 del Anexo.

En **Brasil**, la investigación en PLN es llevada a cabo principalmente en **universidades públicas**, aunque existe alguna actividad en universidades privadas y en empresas privadas. Los grupos de investigación fueron fundados por investigadores que se doctoraron fuera de Brasil e iniciaron los primeros proyectos de PLN. La investigación en PLN en Brasil es variada y se abordan otras lenguas como en inglés y el español.



En Brasil, se estima que hay unos 250 investigadores (incluidos estudiantes de doctorado) con interés en PLN, de los cuales 35-40 son investigadores activos con el PLN como tema principal.

⁵⁴

<http://nlp.lsi.upc.edu/freeling/>

Por su parte, **Portugal creó *Linguateca***, una iniciativa creada oficialmente en 2002 para la creación de un repositorio de recursos y publicaciones sobre el procesamiento del portugués.⁵⁵ Sin embargo, los mejores analizadores sintácticos para el portugués han sido desarrollados por investigadores en Dinamarca y EE. UU. Se estima que un 35% de los investigadores universitarios en Brasil son doctores y un 16% son titulados superiores. Un 62% tienen formación en Informática y en 29% en Lingüística. Un 28% de los investigadores supervisan trabajos doctorales y un 28% participan en proyectos financiados⁵⁶.

Las líneas de investigación en PLN **abarcaban casi todos los ámbitos en los que se trabaja actualmente a nivel internacional**: construcción y uso de recursos léxicos multilingües, adquisición y representación de conocimiento, extracción de información, métodos de aprendizaje automático, traducción automática, resumen de textos, búsqueda de respuesta, sistemas de diálogo, tecnologías del habla, interfaces de lenguaje natural y tecnologías básicas para procesamiento de lenguaje (analizadores morfosintácticos, sintácticos, semánticos, reconocimiento de entidades nombradas, desambiguación de sentidos, etc.) Muchos de los resultados y productos de la investigación se ofrecen con **licencias libres o de acceso gratuito para fines de investigación**. Para el caso del código abierto, se está favoreciendo la creación de comunidades y se está apoyando al sector industrial.

La **investigación competitiva produce diferentes resultados** que pueden variar en función del área de conocimiento correspondiente. Estos resultados se clasifican en: artículos en revistas científicas, tesis doctorales, libros o capítulos, contribuciones a congresos y otras reuniones científicas, documentos e informes y actividades de difusión y comunicación científica. Además de estos indicadores, los sexenios suponen un indicador indirecto de producción científica, y son otorgados por la Comisión Nacional Evaluadora de la Actividad Investigadora (CNEAI).

En 2009 se celebró la **Primera Jornada del Procesamiento Computacional del Catalán**. El evento congregó a 14 grupos y centros de investigación y un total de 166 asistentes⁵⁷. En lo referente a la investigación, se señalaron la duplicación y la dispersión de esfuerzos en la construcción de corpus, herramientas, etc. como principales puntos negativos. De igual forma, se mencionó la **baja interacción entre lingüistas e informáticos** al crear y compartir aplicaciones y recursos, es decir, un déficit de interdisciplinariedad, identificando además una rigidez estructural en la organización docente de la universidad, parcialmente causante de la situación, y la falta de centros de investigación híbridos.

Es preciso hacer notar que **no toda la investigación llevada a cabo por grupos de investigación en España produce o utiliza recursos para el español** o para

El hecho que constituir una sociedad con varias lenguas cooficiales ha permitido el desarrollo de una investigación multilingüe y en especial en el área de la TA, con tecnologías y recursos multilingües, que se ha extendido al ámbito de otras lenguas románicas.



Se considera necesario mejorar la coordinación y fomentar la cooperación entre grupos de investigación y también con las empresas. Estas consideraciones serían también de aplicación al euskera, al gallego y al español.

⁵⁵ <http://www.linguateca.pt/>

⁵⁶ <http://www.ufgd.edu.br/noticias/pne-por-um-brasil-com-mais-investimento-em-pesquisa-e-inovacao>

⁵⁷ P r o c e s a m i e n t o d e l L e n g u a j e
N a t u r a l , n ú m . 4 3 (2 0 0 9) , p p . 3 8 7 - 3 8 8

alguna de las lenguas oficiales. En muchas de las investigaciones llevadas a cabo se utilizan recursos para el inglés por estar estos más desarrollados y por ser estos los que la comunidad internacional prestigia y utiliza como referencia para la compartición de resultados. También se ha producido una diáspora de doctores formados en España, algunos de ellos trabajando actualmente en XEROX, Watson (IBM), Nuance o Google.

6.1.2. Instituciones y entidades de investigación

La Real Academia Española (RAE), junto con la Asociación de Academias de la Lengua Española (ASALE), de los 21 países de habla española, ejerce una **influencia estandarizadora a través de la publicación de varios diccionarios**, los más destacados son el *Diccionario del español* (DiLE), el *Diccionario del estudiante* (DEsen) o el *Diccionario de americanismos* (DAmer). Además de estas obras, la RAE también dispone de un **banco de datos de referencia formado por varios corpus lingüísticos** que contribuyen a la investigación.

Otras Academias, como la Real Academia de Ciencias Exactas, Físicas y Naturales (RACEFYN), la Real Academia Nacional de Medicina (RANM) o la Real Academia de Ingeniería (RAIN) producen **diccionarios de términos de especialidad, vocabularios y repertorios terminológicos** atendiendo a la encomienda estatutaria sobre vigilancia y difusión del uso correcto de los lenguajes de especialidad y que tiene como objetivo último proporcionar, no solo a los profesionales, sino a la sociedad de estos recursos. **Otras lenguas cooficiales como el catalán-valenciano disponen de instituciones propias** como el *Institut d'Estudis Catalans* (IEC) o la *Acadèmia Valenciana de la Llengua* (AVL), el gallego la *Real Academia Galega* (RAG) y el vasco *Euskaltzindia*. Todas estas instituciones elaboran diccionarios, terminologías y corpus. Para el gallego, también el Centro Ramón Piñeiro para la Investigación en Humanidades (CIRP) ha desarrollado importantes recursos para la investigación en PLN. La Academia de la Llingua Asturiana (ALLA) trabaja en diccionarios, toponimia y otros recursos.

El Consejo Superior de Investigaciones Científicas (CSIC) es la mayor institución pública dedicada a la investigación en España y la tercera de Europa. El CSIC tiene como objetivo fundamental el desarrollar y promover investigaciones en beneficio del progreso científico y tecnológico. Sin embargo, el CSIC no incluye en ninguna de sus áreas ni PLN ni la TA. En cambio, sí dispone de algunos recursos terminológicos de interés para el PLN y la TA y mantiene representación en el Comité Técnico de Normalización AEN/CTN 191 de Terminología. **La Institució Catalana de Recerca i Estudis Avançats (ICREA) tiene contratado un investigador en PLN que lidera el grupo TA** de la Universidad Pompeu Fabra en colaboración con otros dos grupos relacionados en la UPF: el TRL-IULATERM y el GLICOM.

Al respecto de instituciones públicas que acogen el PLN como área de interés, es destacable el caso singular **del Reino Unido donde existe el National Centre for Text Mining (NaCTeM)**, el primer centro creado para minería de textos financiado

Existen otras obras como la Ortografía de la lengua española y la Nueva gramática de la lengua española, ambas con una dimensión panhispánica.



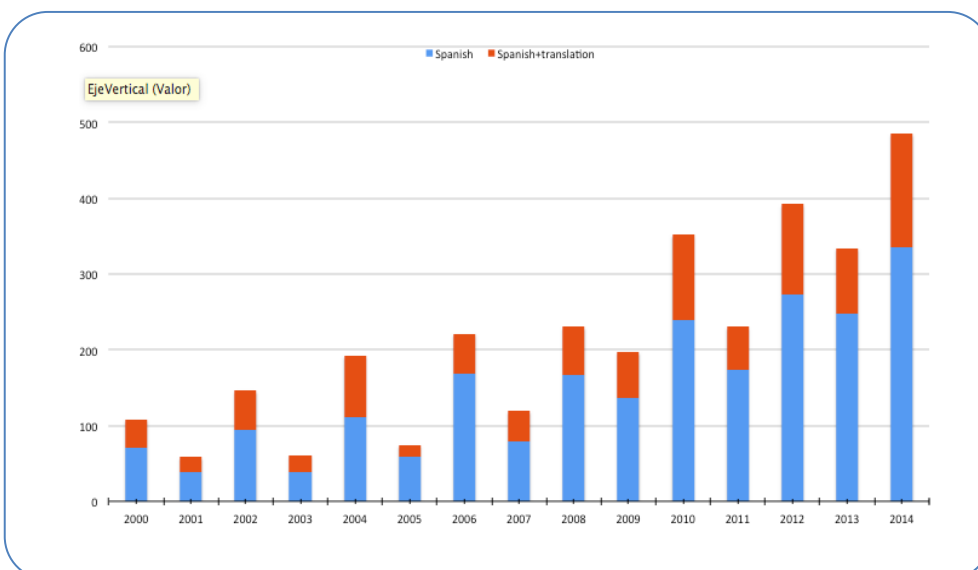
con fondos públicos. Este centro tiene como objetivo proporcionar servicios de minería de textos tanto a la comunidad académica a nivel nacional como a la industria, habiendo tenido gran éxito en el dominio de la biomedicina. Asociaciones nacionales

La **Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)**⁵⁸, es una organización sin ánimo de lucro con más de 320 socios, procedentes tanto del mundo académico como de la industria, que fue creada en 1984 con el objetivo de fomentar y difundir las actividades relacionados con la docencia, la investigación y el desarrollo del PLN, a nivel nacional e internacional. La SEPLN organiza **seminarios, simposios y conferencias y promueve la colaboración entre instituciones nacionales e internacionales**. La presidencia de algunas asociaciones, como la EACL (*European Chapter of the Association for Computational Linguistics*) o la EAMT (*European Association for Machine Translation*), están ocupadas por investigadores españoles. En Latinoamérica cabe destacar la Asociación Mexicana para el Procesamiento del Lenguaje Natural (AMPLN) y el grupo de interés especial en PLN dentro de la *Sociedade Brasileira de Computação* (SBC).

El previamente citado Consorcio MAVIR de la Comunidad de Madrid está formado por siete grupos de investigación de universidades y centros de investigación.

6.1.3. La proyección de la investigación en PLN y TA en español a nivel internacional

En el gráfico que se incluye a continuación se muestra la serie temporal creada ex profeso con el **número de veces que se aparecen el término *Spanish* pero no *translation*** (en azul) y los términos *Spanish* y *translation* (en rojo) en documentos de la Antología del ACL (*Association for Computational Linguistics*), constituida por más de 20 000 artículos, de más de 17 000 autores, y más de 300 conferencias, talleres y números de revistas internacionales.

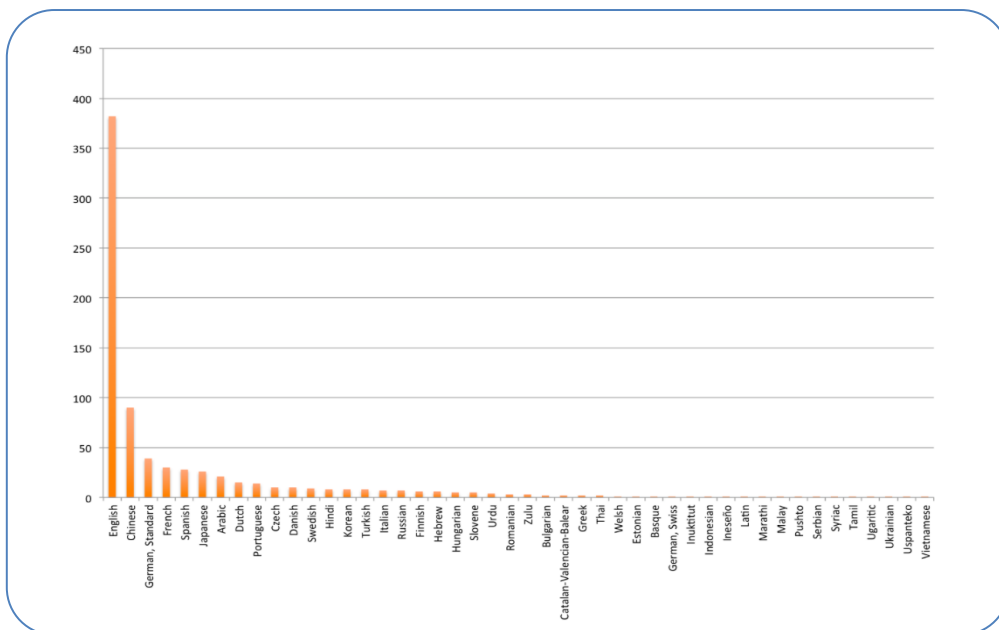


Como puede apreciarse en el gráfico, pese a no ser este un estudio riguroso, el número de publicaciones parece haber aumentado en los últimos catorce años, aunque no de una manera exponencial.

⁵⁸ <http://www.sepln.org/>

Número de documentos que contienen el término *Spanish* pero no *translation* y los términos *Spanish* y *translation* en la Antología del ACL.

Sin embargo, en el gráfico de la figura siguiente publicado por META-NET se muestra la **clasificación de lenguas tratadas en artículos de investigación publicados en revistas y congresos internacionales de referencia de PLN** durante el año 2010. En la gráfica aparecen también el catalán y el vasco. La presencia de recursos textuales en campañas de evaluación internacionales favorecería la investigación en español o lenguas cooficiales.



En el gráfico puede observarse que el español ocupa la cuarta posición a una distancia considerable de las primeras posiciones, ocupadas por el inglés y el chino.

Lenguas tratadas en trabajos de investigación publicados en el *Journal of Computational Linguistics* y las conferencias del ACL, EMNLP y COLING. (fuente: META-NET, *Strategic Research Agenda for Multilingual Europe 2020*, Versión 0.9)

6.2. Principales fuentes de financiación de la investigación española en PLN y TA. Ayudas públicas y privadas en España y Europa

6.2.1. Instrumentos y estrategias de financiación para la investigación en PLN y TA

El informe de la CRUE (Conferencia de Rectores de las Universidades Españolas)⁵⁹ Investigación y Transferencia 2012 y 2013 de las universidades españolas, elaborado por Grupo de Trabajo de Indicadores, RedOTRI y RedUGI, y publicado en Junio 2015, ya constata que en **un 80% los fondos con los que se financia la investigación en las universidades en España es fundamentalmente de origen público** y esta área sigue la misma tendencia, sea con financiación interna de las universidades, básicamente PDI propio, o con financiación externa obtenida en convocatorias competitivas. Las convocatorias



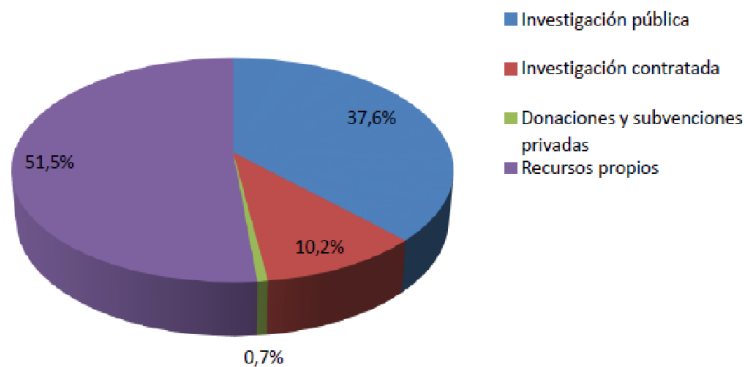
59

http://www.crue.org/Documents/Informe%20Encuesta%20ITC%20RedOTRI_2012y2013.pdf

competitivas son en su mayoría la forma de acceder a fondos de origen estatal (Plan Estatal), de comunidades autónomas o entes locales y de ámbito europeo (Programa Marco) e internacional (NSF).

La investigación en Procesamiento del Lenguaje Natural en España se lleva a cabo principalmente en departamentos de universidades, como se ha podido ver previamente.

Figura 2.4. Perfil de la I+D universitaria según el origen de los fondos (porcentaje de financiación). Anualidad 2013.



Fuente: Encuesta I+TC 2013 (60 respuestas válidas).

La financiación obtenida en convocatorias competitivas permite **sufragar los siguientes costes de la investigación:**

- Contratación directa de personal de apoyo a la investigación y técnicos
- Financiación para la formación de personal investigador que actualmente está ligado al desarrollo de proyectos de investigación financiados,
- Financiación de integración de investigadores doctores.
- Adquisición y mantenimiento de equipamiento científico
- Movilidad, difusión y publicación del conocimiento

La principal fuente de financiación es el Plan Nacional de I+D o el vigente actualmente "Plan Estatal de Investigación Científica, Técnica y de Innovación 2013-2016", gestionado por la Secretaría de Estado en I+D+i (SEIDI), que tiene los siguientes ejes y acciones prioritarias, y donde no hay ninguna referencia destacada al PLN y TA.

- **Excelencia:** investigación fundamental, no orientada.
- **Retos :** proyectos de investigación orientada a dar respuesta a los retos de las sociedad⁶⁰:
 - Salud, cambio demográfico y bienestar.



⁶⁰ [Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016](#), a partir de la página 27

- 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 cambio climático y eficiencia en la utilización de recursos y materias primas.
- Cambios e innovaciones sociales.
- Economía y sociedad digital.
- Seguridad, protección y defensa.

- **Acciones estratégicas.**

- Ae1. Acción estratégica en salud.
- Ae2. Acción estratégica en economía y sociedad digital.

Además de la financiación pública, se observa en los datos suministrados por la CRUE un número, aún pequeño pero creciente, de **convocatorias competitivas con financiación privada** (BBVA, RecerCaixa, son algunos ejemplos), pero tradicionalmente la financiación privada se materializa en la contratación de servicios y/o desarrollo por parte de empresas interesadas en nuevas soluciones.

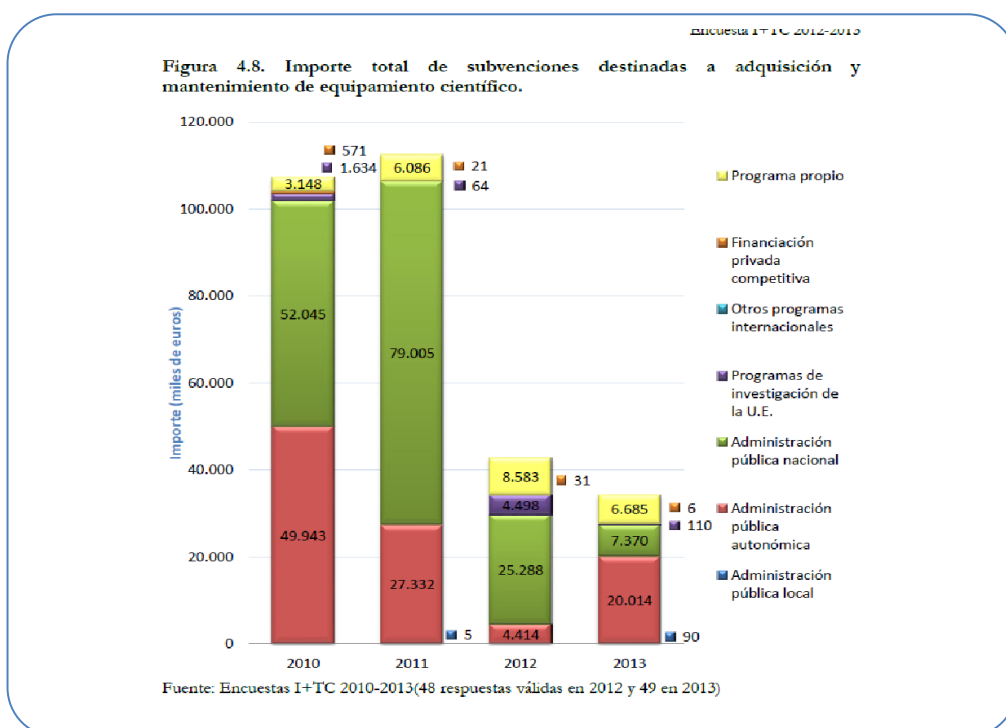
Hay que resaltar que la financiación competitiva pública de los grupos de investigación universitarios también **incluye acceder a esquemas mixtos de co-financiación empresa-subservención** para colaboraciones de investigación, que se explica más abajo. Actualmente, en el programa RETOS del Plan Estatal se propone un esquema de colaboración público-privada: RETOS-Colaboración.

En la financiación para la formación de investigadores (Formación de Investigadores FI, Formación de profesorado Universitario, FPU) se establecen además algunos **esquemas mixtos en los llamados "doctorados industriales"** que complementan ayudas mixtas anteriores que apoyaban la integración de doctores en la empresa (Programa Torres Quevedo, por ejemplo). Otra línea de financiación es la del Plan Estatal de Promoción del Talento y su Empleabilidad con los programas Ramón y Cajal, Juan de la Cierva con los que incorporar al grupo investigadores post-doctorales en diferentes fases de la carrera investigadora.



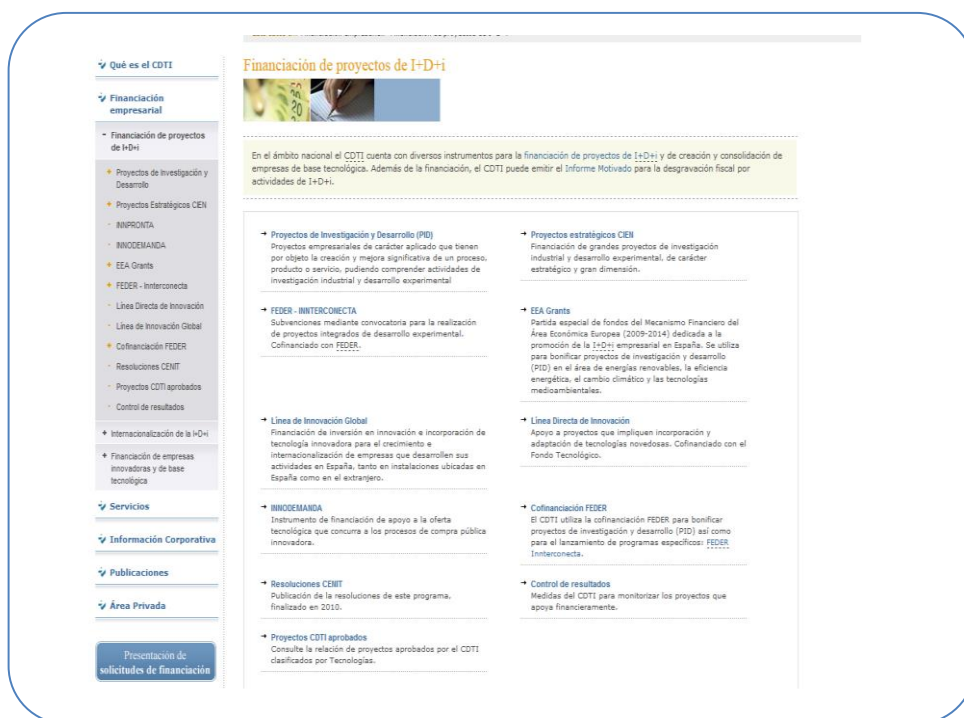
La cofinanciación pública-privada también puede obtenerse gracias a convocatorias específicas, en su mayoría gestionadas por el Centro de Desarrollo Tecnológico Industrial (CDTI), u organismos análogos en las comunidades autónomas, y en algunos casos por organismos ministeriales con objetivos de desarrollo concretos: SETSI, SEGITTUR, etc.

Hay una **amplia variedad de fuentes de financiación**, de naturaleza muy variada, aunque normalmente tienen esquemas de financiación parcial (no siempre es subvención directa del 100% de los costes marginales de la investigación incluyendo contratación de personal) lo cual incide en la adecuación para proyectos de naturaleza específica, como son los de Procesamiento del Lenguaje Natural y Traducción Automática. Estos se caracterizan por la **necesidad de contratar programadores informáticos o investigadores de lingüística computacional**, compra de infraestructura informática o de servicios de computación y gastos de difusión (un área caracterizada por el mayor impacto de las publicaciones en congresos internacionales que por las revistas, por ejemplo, solamente tiene 3 revistas indexadas).



La financiación de la investigación en las empresas tiene como eje principal el Centro de Desarrollo Tecnológico Industrial (CDTI). Ofrece ayudas en forma de subvenciones y préstamos blandos retornables a empresas de todo el rango.

La **financiación pública competitiva con fondos del programa marco de la Comisión Europea** es la otra gran fuente de financiación para los proyectos de investigación de los grupos españoles. De más reciente creación son las convocatorias internacionales conjuntas, como el programa CHIST-ERA.



Áreas de financiación de I+D+i del CDTI. Captura de pantalla, 9 julio 2015. www.cdti.es

6.2.2. Eficacia y financiación obtenida

En general, las fuentes de financiación de los grupos de investigación han sido las **convocatorias competitivas a nivel regional, nacional y europeo ya mencionadas**, y, en algún caso, vía la **Agencia Española de Cooperación Internacional (AECID)**. El Ministerio de Educación, a través de la Comisión Interministerial de Ciencia y Tecnología (ICYT) apoya la investigación en el campo de las tecnologías de la información, a través de programas nacionales de investigación. La base del desarrollo tecnológico y de las aplicaciones comerciales para el procesamiento automático del español ha surgido en parte como resultado de estos proyectos.

También el **Centro para el Desarrollo Tecnológico Industrial (CDTI)**, dependiente del Ministerio de Ciencia e Innovación, a través de distintos programas, contribuye a la financiación de proyectos de I+D de empresas españolas que quieren mejorar su perfil tecnológico. Los proyectos financiados por el CDTI en el área de Tecnologías de la Información y Comunicación a menudo incluyen algún tipo de **investigación en tecnologías del lenguaje y participan en ella centros públicos de investigación y centros tecnológicos**.

Los **grupos de investigación españoles tienen participación en el programa ICT de Tecnologías de la información y la comunicación, bajo el Séptimo Programa Marco de la UE (FP7)**. Según información recopilada por Zabala Innovation Consulting, estos grupos de investigación participan en más de una veintena del centenar de proyectos de investigación del programa, y la

Estos programas han impulsado numerosos proyectos de investigación y de colaboración con empresas y centros internacionales de investigación.



financiación de estos proyectos se estima que alcanza los 65 millones de euros. Salvo la consultora Zabala Innovation Consulting, que coordina el proyecto LT Observatory y que tiene entre sus objetivos la identificación de recursos lingüísticos y de TA en Europa, solo hay un grupo español actualmente participando en proyectos financiados en el programa Horizonte 2020, el proyecto KRISTINA, con un consorcio de nueve grupos/empresas y una financiación de 3 633.801 euros.

El **Gobierno Vasco da soporte a la investigación e innovación a través del Plan de Ciencia y Tecnología (PCTI)**. En todos los planes, las tecnologías del lenguaje han sido identificadas como área estratégica. La mayoría de los recursos y herramientas para el vasco se han obtenido a través de proyectos financiados por estos planes durante los 10 últimos años. La **Xunta de Galicia apoya la investigación a través del Plan Galego de Investigación, Desenvolvemento e Innovación Tecnolóxica** (PGIDIT). A pesar de que la tecnología lingüística no es un área prioritaria, a lo largo de los años, los grupos de investigación de las universidades y alguna empresa tecnológica han conseguido bolsas para llevar a cabo investigaciones y desarrollos en tecnologías del lenguaje.

En **Brasil, la financiación en la actualidad proviene principalmente de cuatro agencias gubernamentales**. La financiación privada todavía es escasa, lo cual refleja la **poca interacción entre universidad y empresa**. Es necesario hacer notar que las lenguas co-oficiales, al no ser oficiales a nivel europeo (hay 24 lenguas oficiales) no reciben ningún tipo de ayudas específicas a nivel europeo. La incorporación de soporte a estas lenguas en proyectos europeos se hace por iniciativa propia. La comunidad investigadora en Cataluña ha recibido soporte en los programas de investigación español y catalán. En el pasado existieron centros como el CREL (*Centre de Referència en Enginyeria Lingüística*, 1996-2000), mencionado anteriormente. Más recientemente, desde 2008, **el catalán ha tenido presencia en proyectos europeos de infraestructuras como CLARIN**, que también ha contado con apoyo de FEDER, y METANET4U (relacionado con META-SHARE con participación del grupo TRL-IULATERM) que identifican y dan acceso a recursos estándar y tecnologías para PLN y TA.

En el informe de la CRUE se advierte ya **un crecimiento del programa marco de la UE como fuente de financiación**, aumento que también se había cumplido en el caso de estudio en el transcurso del VII Programa Marco. Los grupos españoles en PLN y TA acceden a fondos europeos básicamente en el área de Tecnologías de la Información y las Comunicaciones (TIC). Dentro de este ámbito, y a lo largo de sus diferentes ediciones, se ha hecho patente la importancia del multilingüismo y del procesamiento del lenguaje natural presente desde el IV al VII Programa Marco, con una línea específica para "Tecnologías Lingüísticas", que abarcaba PLN y TA, reconocimiento y síntesis del habla y en el que los grupos españoles obtuvieron resultados que reflejan la masa crítica de esta área actualmente. Varios informes como el elaborado por META-NET detectan **una falta de continuidad en la financiación de la investigación y el desarrollo**, alternándose programas coordinados a corto plazo con períodos de escasa o nula financiación. Además,

El número de participantes de la veintena de proyectos es de 172 y reciben una media de 380 000 euros por grupo (con lo que los grupos españoles recibirían cerca de 7-8 millones de euros en total).



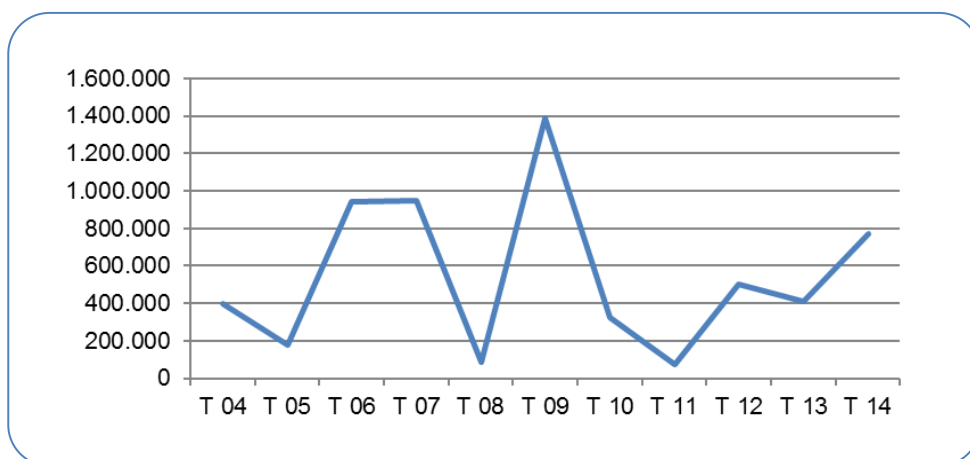
parece haber una falta general de coordinación con los programas de otros países de la UE y de la Comisión Europea. A pesar de la interrupción o la reducción en España de las actividades en tecnologías del lenguaje por parte de las grandes empresas, existe una pequeña industria y un marco de investigación que en el pasado se benefició de programas de investigación importantes que han ayudado a la creación de tecnologías y recursos que todavía son limitados y están lejos de los que disponen lenguas como el inglés.

En el Plan Nacional, el PLN y la TA tuvieron una priorización en el de 2004-2007 y en el de 2008-2011 en los **siguientes términos**:

El programa TIN incluyó como acción estratégica **la ingeniería del lenguaje humano adaptado al castellano y otras lenguas del Estado español** para la **mejora de la base tecnológica que soporta el reconocimiento y procesamiento del castellano y otras lenguas oficiales** del estado español.

Las actividades propuestas abarcaron desde la **investigación básica en áreas como el aprendizaje automático, la lingüística computacional** (análisis sintáctico y expansión semántica, entre otros), o el **diseño de motores de búsqueda** basados en el procesamiento del lenguaje natural, a temas más aplicados como el **trabajo colaborativo** y otros cercanos al área de humanidades (comunicación multicultural, etc.). Además se planteó que la acción se realizará en común con el Área de Humanidades, Ciencias Sociales y Económicas, y tuviera la participación de otros agentes nacionales (Real Academia de la Lengua, Instituto Cervantes, etc.).⁶¹

La financiación obtenida por los grupos de investigación universitarios en estos mismos períodos del Plan Nacional y Plan Estatal se ve en la siguiente tabla.



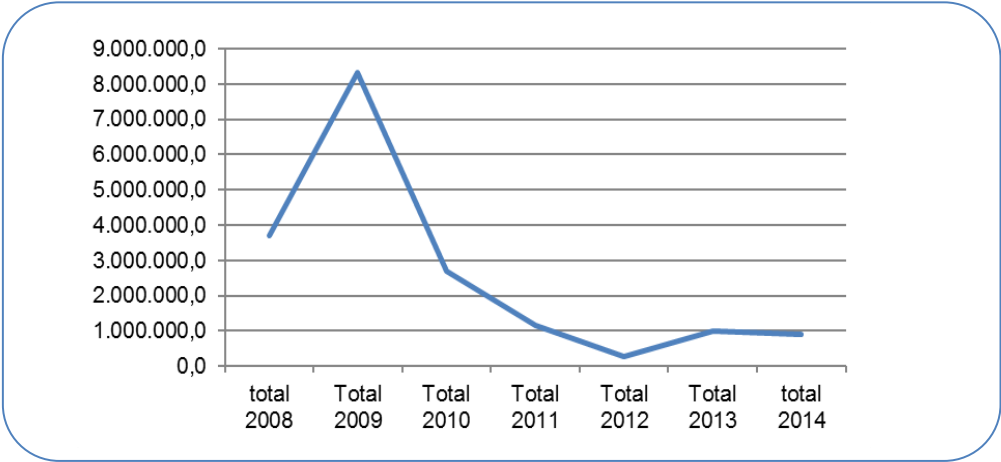
Financiación obtenida por los grupos de investigación españoles en el Plan Nacional/Estatal: 12,622 M€ en el período 2004-2014. Financiación media por proyecto de tres años: 170.000€

Señalar que en el actual Plan H2020 el apoyo al PLN y TA se ha reducido considerablemente.



⁶¹ Plan Nacional 2008-2011

Además de la financiación CDTI, otras entidades, por ejemplo la Secretaria de Estado de Telecomunicaciones y Sociedad de la Información (SETSI)⁶² tiene también un **programa de apoyo a la innovación en empresas** que han sido **una importante fuente de financiación para las empresas del sector**. En el siguiente gráfico se ve la evolución de la financiación de proyectos del sector en el período 2008-2014.



Por último, en el siguiente informe se referencian los fondos que grupos españoles (universidad y empresa) han conseguido en las convocatorias del 7PM.



PROGRAMA MARCO

Control de Retornos

Fecha Impresión: 07/07/2015

Descripción:. Programa Marco: 7PM; Programas: 7ICT; Áreas: 04.03 - (7ICT) TECNOLOGÍAS LINGÜÍSTICAS.

I.-PROPUESTAS:

	Esp	Total	%Esp/Total
Nº Propuestas	89	212	42,0
Nº Socios	77	770	10,0
Nº Líderes	29	212	13,7
Nº Participaciones	151	1.502	10,1
Presupuesto	69.269.318	783.850.664	8,8
Subv. solicitada(Euros)	53.203.771	590.769.137	9,0

II.-ACTIVIDADES FINANCIADAS:

	Esp	Total	%Esp/Total
Nº Propuestas	24	44	54,5
Nº Socios	23	194	11,9
Nº Líderes	9	44	20,5
Nº Participaciones	39	299	13,0
Presupuesto	20.740.479	157.119.097	13,2
Subv. solicitada(Euros)	15.608.399	119.130.551	13,1

13,76% UE

	Esp	Total
Subv. Media por Participantes (Euro)	400.215	398.430
Subv. Media por Coordinador (Euro)	578.463	710.205
Subv. Media por Actividad (Euro)	650.350	2.707.513
Promedio de participaciones por Actividad (nº)	1,6	6,8
Propuestas aprobadas/presentadas (%)	27,0%	20,8%

62 <http://www.ipyme.org/es-ES/BBDD/Paginas/HomeBBDD.aspx>

Financiación de grupos españoles en el 7PM. Fuente CDTI. Período 2007-2013

6.3. Análisis DAFO

Oportunidades

- Siendo el español una **lengua transnacional de casi 500 millones de hablantes** existe una oportunidad de investigación y de mercado con una competencia a nivel internacional.
- Existe una iniciativa del Gobierno español para el **mantenimiento de un catálogo de datos abiertos**, en el que se pueden incluir datos para PLN o TA.

Fortalezas

- El **modelo plurilingüe** español favorece especialmente la investigación en PLN y TA.
- Los **grupos de investigación universitarios españoles** están consolidados y permanecen estables según puede apreciarse en los censos realizados a lo largo de los últimos 15 años. Hay censados en España más de 30 grupos de investigación y más de 160 investigadores, principalmente en universidades⁶³. Existen asociaciones y redes temáticas sobre PLN en español formadas principalmente por grupos de investigación universitarios. Las presidencias de las asociaciones europeas: *European Association for Computational Linguistics*⁶⁴ y *European Association for Machine Translation*⁶⁵ están ocupadas por investigadores españoles.
- Las **líneas de investigación en PLN en España** abarcan casi todos los ámbitos en los que se trabaja actualmente a nivel internacional.

Debilidades

- Los grupos de investigación en PLN y TA en grandes empresas en España han **prácticamente desaparecido** (IBM, TID, etc.).
- Las Agencias Públicas de Investigación (p.e. CSIC) no contemplan en ninguna de sus áreas científico-técnicas **ningún tipo de investigación básica ni desarrollo tecnológico** en torno al PLN o la TA. La ICREA tiene contratado 1 investigador en PLN⁶⁶.
- Se percibe **duplicación y dispersión de esfuerzos** en la construcción de corpus, herramientas, etc. No hay líneas de trabajo consolidadas entre los grupos de investigación en PLN e instituciones como las Reales Academias o similares, tampoco entre la Reales Academias



⁶³ www.sepln.org
⁶⁴ <http://www.eacl.org/page.php?id=officers>
⁶⁵ <http://www.eamt.org/committee.php>
⁶⁶ lcrea.cat

- Hay un **déficit de interdisciplinariedad** debido a la baja interacción entre lingüistas e informáticos en la creación y compartición de recursos y aplicaciones y motivado por la rigidez estructural universitaria y la falta de centros de investigación sobre PLN y TA.
- La investigación y desarrollos realizados en lenguas co-oficiales están **lejos del español**, y el español está lejos de la investigación llevada a cabo para el inglés⁶⁷.
- Se han **generado algunos recursos para las lenguas del estado** de libre disposición o código abierto que apoyan al sector industrial.
- Algunos **grupos de investigación participan en consorcios industriales** tanto nacionales como en el Programa Marco de la UE.
- Las instituciones relacionadas con las lenguas españolas (RAE, IEC, etc.) han ejercido y siguen ejerciendo **una labor reguladora y estandarizadora** sobre el lenguaje, tienen una dimensión panhispánica, forman parte de redes con Iberoamérica, y gozan de prestigio en la sociedad.

Amenazas

- **Grandes compañías como Google** compiten en investigación y desarrollo con los grupos de investigación de PLN y TA tanto en español como en lenguas cooficiales.
- En los informes de META-NET se **detecta falta de continuidad en la financiación de la investigación y el desarrollo**. Se alternan programas coordinados a corto plazo con períodos de escasa o nula financiación. Parece haber una falta general de coordinación con programas de otros países de la UE y de la Comisión Europea.



⁶⁷

<http://www.meta-net.eu/whitepapers/overview>



7.

Situación del Mercado y la
Industria del lenguaje en
español y lenguas co-
oficiales

7. Situación del Mercado y la Industria del lenguaje en español y lenguas co-oficiales

7.1. Estudio del sector industrial del PLN en español y lenguas cooficiales

La información de este apartado ha sido obtenida a través de la **elaboración de una encuesta que ha sido enviada a un total de 34 empresas españolas**. Todas ellas consideradas como empresas con tecnología en el ámbito de la TA o bien del PLN. No se han considerado por tanto otras empresas que son usuarias de forma activa de estas tecnologías, como pueden ser empresas de traducción, editoriales, gestoras y generadoras de contenidos, aprendizaje de lenguas, etc.

Las **empresas objetivo** han sido identificadas a través la información suministrada en los siguientes grupos y organizaciones:

- **Langune** (organización cluster de empresas de PLN y TA en el País Vasco)
- **Clusterlingua** (organización cluster de empresas de PLN y TA en Cataluña)
- **LT-Innovate** (organización de empresas de PLN y TA en Europa)
- **TIMM** (Red Temática en Tratamiento de Información Multilingüe y Multimodal)
- **Consorcio MAVIR** (red de investigación co-financiada por la Comunidad de Madrid dentro del IV Plan Regional de Investigación Científica e Innovación Tecnológica (IV PRICIT) y formada por un equipo multidisciplinar de científicos, técnicos, lingüistas y documentalistas para desarrollar un esfuerzo integrador en las áreas de investigación, formación y transferencia de tecnología)

La encuesta tiene como objetivo **obtener información relativa** a:

- Área de las industrias de la lengua en las que trabaja

La encuesta, convenientemente adaptada, se envió también a otras tantas empresas europeas para obtener una foto de las características de las empresas, también en el ámbito europeo con las que poder compararse.



- Ámbito geográfico de su mercado
- Lenguas principales en las que ofrece sus servicios
- Tipo de oferta de servicios/productos
- Aspectos relativos a la tecnología (propiedad, especialización, etc.)
- Recursos o herramientas que precisan en su actividad habitual
- Aspectos sobre la estructura de la empresa (empleados, relación con universidades, etc.)
- Aspectos relativos al I+D: líneas de investigación, financiación, patentes, etc.
- Volumen de facturación
- Necesidades más importantes



7.1.1. Resultado de la encuesta (ámbito español)

La industria de la lengua, visto como sector más amplio, engloba **otras áreas con un peso económico mucho mayor**: traducción, editoriales, enseñanza de idiomas, grupos de comunicación multilingüe, turismo, etc. No obstante, tal y como se ha indicado anteriormente, la encuesta no estaba dirigida al sector 'ampliado'.

Los **resultados más relevantes** de la encuesta se muestran a continuación:

PLN/TA

- 7 empresas en el área de PLN
- 8 empresas en el área de TA

Lenguas de trabajo

- Cubren mayoritariamente las principales lenguas europeas y en gran medida, también, las lenguas cooficiales.

Tamaño de la empresa

- Se trata de PYMES, de entre 10 empleados mínimo (con alguna excepción) y 50-60 máximo.

Propiedad de la tecnología

- En su mayoría las empresas consultadas son propietarias de su tecnología. Solamente 3 de las empresas consultadas distribuyen la tecnología bajo licencia.

Especialización de la tecnología y variantes lingüísticas

- En la mayoría de casos la tecnología es válida para cualquier dominio y permite cierta personalización o adaptación.

La encuesta ha sido respondida por 14 empresas de las 34 identificadas.

- Se cubren las variantes lingüísticas del inglés, portugués, catalán, etc.

Relación con la universidad

- Las empresas mantienen programas de colaboración con universidades pero pocas de ellas están implicadas en la dirección de tesis o en la impartición de postgrados o masters.

I+D+i

- Algunas de las empresas (6) han realizado publicaciones científicas.
- Solamente una de las empresas ha desarrollado una patente.
- Los proyectos de investigación son mayoritariamente de ámbito nacional y/o europeo.
- Algunas de las empresas (6) disponen de líneas de financiación públicas.
- Sobre las líneas de investigación principal: hibridación de STA, sistemas de TA en la nube, entornos de postedición, paralelización, extracción de relaciones, análisis del sentimiento.



Sobre el perfil de los colaboradores

- Se constata que existen muy pocos doctores en las plantillas de las empresas, aunque las empresas manifiestan que disponen de personal especializado en el área.

Volumen de facturación

- Entre 200.000 y 4.000.000 de €

Modelo de negocio

La mayoría de empresas del sector **orientan su actividad en el área de servicios**, desarrollo de recursos y desarrollo de productos. Las desarrolladoras ofrecen también servicios de mantenimiento y personalización de sus soluciones.

Las herramientas, aunque son válidas para cualquier dominio, precisan de una **adaptación y personalización**. Esta personalización requiere, por una parte, disponer de máxima cooperación por parte del cliente para que aporte recursos (glosarios, diccionarios, textos paralelizados,) y, por otra, requiere que las empresas tecnológicas dispongan de mecanismos para obtener recursos básicos de calidad para proceder a esta personalización. Este es quizás uno de los puntos donde se ve que las empresas pueden tener dificultades para poder ofrecer de una forma rápida, acotable y evaluable a priori, procesos de personalización donde intervengan recursos básicos adquiridos fuera de su entorno empresarial: resulta difícil conseguir recursos que sean compatibles y que el proceso de incorporación a la tecnología sea sencillo. Finalmente, **evaluar a priori el impacto de esta mejora** resulta también una tarea muy difícil ya que cada proyecto es diferente y las variables que lo configuran son distintas (área, tipología de los textos, lengua, entorno...).

Debido a que las TL interactúan cada vez más con otras tecnologías y procesos, tiende a desaparecer la modalidad de comercialización de producto cerrado y resulta cada vez más común la comercialización de las soluciones a través de API (Application Programming Interface). De esta manera la funcionalidad deseada se integra fácilmente en entornos más complejos.

Dificultades para llegar al mercado

Las principales **dificultades** para llegar al mercado son:

- **Visibilidad:** muchos proyectos que utilizan tecnologías lingüísticas ‘ocultan’ este hecho ya que a menudo puede percibirse este hecho como algo que va en detrimento de la calidad lingüística o que supedita una lengua a un nivel inferior respecto a otra.
- Las **expectativas** de los clientes suelen ser muy altas. Se suele esperar que con el uso de estas tecnologías el problema lingüístico se solucione por completo.
- **Falta de metodología/métricas** claras que ayuden a planificar correctamente los proyectos de forma que las curvas de aprendizaje y los resultados (ahorros) obtenidos puedan ser los esperables a priori.
- Hace falta avanzar también en la **definición y tipologías de proyectos** concretos, en función de los resultados esperables
- Existe todavía un cierto **recelo a suministrar** datos masivamente para poder proceder a personalizar y mejorar los sistemas.
- Algunas **tecnologías lingüística** tienen limitaciones intrínsecas para tratar nuevas lenguas o bien para poder diferenciarse de forma clara respecto otras tecnologías.

En el caso de la TA se perciben dificultades claras en el encaje entre la oferta de las empresas y la demanda del mercado. Algunas posibles explicaciones de dicho desencaje se describen a continuación.

Sobre la tecnología

- La tecnología de TA, como producto que ‘soluciona completamente’ el problema de la **traslación de información entre lenguas** se percibe como una tecnología que provoca frustración en las expectativas del cliente.
- Se trata además de una tecnología con cierto recorrido histórico. Las primeras aproximaciones tecnológicas a la resolución del problema de la traducción se remontan a mediados del siglo XX. Aunque la tecnología ha avanzado mucho y la capacidad de computación de los ordenadores ha aumentado exponencialmente, **el problema sigue siendo percibido como ‘no resuelto’** (al menos de forma completa).
- La utilización de la TA de forma productiva (es decir, de forma que pueda competir en costes y resultados con procesos de traducción convencional) requiere que el proyecto permita la definición de fases previas de preparación de los motores de traducción y que los procesos de postedición se definan en términos que favorezcan la retroalimentación del sistema de traducción automática.
- La **metodología para la implantación de TA en un proceso productivo**, aunque se base siempre en un proceso previo de aprendizaje y una posterior retroalimentación a partir de la experiencia del proceso de postedición, resulta de muy difícil evaluación ya que cada proyecto es diferente (tipología de textos, calidad de los contenidos a traducir, área



temática, plazos requeridos para la publicación, ...) y por lo tanto la evaluación previa de la fase de personalización de los motores y la evaluación de la curva de mejora/aprendizaje que se obtendrá en la fase de postedición resultan muy difíciles de acotar. A todo ello hay que sumar que al final, el verdadero proceso de mejora/aprendizaje viene dado por la adecuación de la destreza y perfil del posteditor humano con las limitaciones del motor de traducción.

- Existen **pocas métricas claras que permitan extraer y extrapolar consecuencias de las curvas de aprendizaje**. Por supuesto que en los casos de éxito siempre se pueden establecer medidas en términos de ahorros conseguidos (respecto a procesos de traducción convencional) pero resulta prácticamente imposible extrapolar resultados a otros proyectos o establecer objetivos tendencia de mejoras esperadas para las etapas siguientes del proceso.
- Las demandas del mercado requieren a menudo plazo de ejecución en los proyectos tan cortos que **imposibilitan la definición correcta de un proyecto de forma adecuada**, con las fases de aprendizaje y mejora del sistema que se requieren. La ejecución a corto plazo imposibilita también la generación de recursos básicos como extracción de terminología, generación de memorias de traducción, elaboración de glosarios, etc.
- A menudo la tecnología de TA se utiliza en proyectos donde **el cliente no quiere visibilizar su uso**. Algunos de los casos de éxito más claros y contundentes de nuestro entorno lo demuestran.
- Se requieren **recursos básicos que no siempre están disponibles**: textos traducidos en ámbitos lingüísticos particulares, diccionarios, glosarios, textos anotados, listados de nombres propios. Además para poder agilizar los procesos de personalización conviene que estos recursos sean abiertos, libres o como mínimo exista facilidad de interoperabilidad entre distintos formatos.

Sobre las limitaciones de las empresas

- Los proyectos susceptibles de aplicación de la tecnología de TA suelen requerir, además de **la solución del problema de traducción, la resolución o gestión de problemáticas** que pueden ser más complejas que el propio problema de la traducción: gestión de formatos de la información; extracción, manipulación y publicación de contenidos en herramientas de gestión complejas; control de flujos de publicación con exigencias de inmediatez o sincronía entre las diversas lenguas; problemas de localización de los contenidos (a menudo, al plantearse el multilingüismo en una organización surgen cuestiones que no habían sido consideradas en un entorno monolingüe), procesos de producción editorial donde todavía prevalecen tareas con niveles de automatización muy bajo, etc.
- Las empresas, por lo tanto, son **capaces de configurar una oferta muy convincente y potente** en lo que se refiere al ámbito de la de TA pero no



Estas circunstancias hacen que a menudo las empresas de tecnología de TA se vean obligadas a ser conocedoras también de tecnologías colaterales en las que claramente no pueden ser tan competitivas (gestores de publicación, gestores de contenidos, organización de procesos de publicación, entornos de publicación multilingüe en tiempo real, etc...).

son capaces (o no tienen suficiente competencia técnica para ello), de influir en el resto del conjunto del proyecto.

- Existen proyectos donde **no confluyen las complejidades de otras tecnologías ni problemáticas**. En estos casos suelen primar parámetros para la obtención de resultados económicos satisfactorios que no permiten configurar el proyecto en términos de mejoras y retroalimentación de la tecnología de TA.
- Las empresas están ejecutando las tareas **de post-edición con personas que profesionalmente han sido formadas en el área de traducción**. Se considera prioritario definir claramente el ámbito de la post-edición (donde empieza, donde acaba, que calidad es esperable en un proceso de TA+PE) y la formación específica de los profesionales de post-edición.
- El **tamaño de las empresas** condiciona en muchas ocasiones las acciones y el nivel de proyectos que pueden acometer.

Se constata, además, que el grupo de empresas del sector es muy reducido si se toman como referencia solamente aquellas que producen o comercializan tecnología. Se constata también que existen bastantes centros de investigación en el ámbito universitario. Centros que producen recursos o tecnología básica y que juegan un papel importante en el sector.

7.1.2. Resultado de la encuesta internacional (Europa)

Se lanzó una encuesta a empresas extranjeras (correos extraídos de los miembros de LT-Innovate) se obtuvo respuesta de 20 organizaciones. **No se ha podido obtener más información** que las que se desprenden de los resultados de la encuesta puesto que finalmente el informe de LT-Innovate sobre el mercado latinoamericano no ha podido obtenerse. Los resultados de la encuesta en su totalidad pueden consultarse en los anexos.

Como **cuestiones a destacar**, que marcan claramente puntos diferentes de las empresas españolas, cabe mencionar:

- La mayoría de las empresas indican que **su área de trabajo es PLN**. Solamente una indica TA. Ocho empresas indican ambas áreas de trabajo.
- Las **lenguas de trabajo son básicamente las europeas**. Algunas empresas incluyen también lenguas asiáticas.
- En relación al tamaño, las respuestas indican que los tamaños de las empresas continúan siendo pequeños, dentro de las proporciones de las PYMEs. Destaca solamente una empresa (Nuance) con un tamaño excepcionalmente mayor a la media (+10000)
- En los perfiles de los colaboradores se observa **mayor presencia de doctores**.
- Las **cifras de facturación son similares** a las que indican las españolas, aunque muchas empresas han evitado dar respuesta a esta pregunta.



7.1.3. Casos de uso

La existencia de diversas lenguas cooficiales en España ha sido sin duda un **factor clave en el desarrollo de las tecnologías de la lengua**. En los últimos años se ha visto cómo estas tecnologías han sido determinantes en el éxito de algunos proyectos que sin su participación hubiese sido imposible ejecutar.

Prensa escrita

Desde 1998 años la tecnología de traducción automática está presente en el proceso productivo de la edición catalana del periódico Segre (Lleida). Después siguieron esta iniciativa otras cabeceras como El Periódico de Catalunya en 1999 y desde hace 4 años lo hizo un diario centenario como es La Vanguardia. En todos estos proyectos **la integración de una herramienta de traducción automática en el proceso editorial constituye la pieza fundamental** del proyecto. Otros aspectos clave de este tipo de proyectos son la definición del proceso de post-edición y la permanente retroalimentación del motor de traducción.

Administración pública

Son muchas las **administraciones públicas que disponen de sus herramientas de traducción automáticas**. El nivel de integración varía según sea la funcionalidad que se le ha querido otorgar al sistema de traducción automática. Así pues casi todas las administraciones autonómicas utilizan traductores automáticos accesibles desde su intranet para la traducción de su documentación (Generalitat de Catalunya, Xunta de Galicia, Gobierno Vasco, Gobierno de Baleares). Algunas de ellas tienen integradas las herramientas de TA en los gestores de contenidos web para facilitar la edición y mantenimiento de sus webs en diferentes lenguas (por ejemplo, el Gobierno de Baleares).

Otras aplicaciones dentro de la administración pública son: la traducción del BOE a las lenguas cooficiales y los portales de herramientas de traducción de uso abierto a la ciudadanía como el utilizado por la Dirección de Tecnologías de la Información y las Comunicaciones dependiente del Ministerio de Hacienda y Administraciones. Públicas basado en PLATA.

Entornos editoriales

Los traductores automáticos también están presentes en los **flujos de producción de contenidos multilingües de las editoriales de libros de texto** (Anaya educación, Xerais, Algaída) Los altos niveles de productividad que pueden alcanzarse entre lenguas próximas como el castellano, catalán (con sus diferentes variantes) y gallego hacen de la TA una herramienta imprescindible en la generación de contenidos de los libros de texto.

En el ámbito digital existen actualmente muchas cabeceras de periódicos que realizan traducciones a diversas lenguas pero resultan destacables (también por el uso de TA a otras lenguas distintas de las cooficiales) la versión que el diario El País hace de su web para Brasil y la versión catalana para los lectores catalanes, ambas con TA.



Atención al cliente

La **empresa catalana Inbenta** ha conseguido una amplia implantación de sus sistemas de búsqueda semántica en entornos de atención al cliente. Mediante la utilización de procesamiento de Lenguaje Natural y técnicas de inteligencia artificial optimiza los flujos de respuesta a los requerimientos de los clientes. Actualmente se utiliza este tipo de soluciones en grandes empresas de todos los sectores.

Gestión de la información

Daedalus aplica su tecnología de procesamiento de lenguaje para conseguir **soluciones eficientes en los campos de búsqueda de información, tratamiento automatizado de los procesos de corrección lingüística** y también en el área de inteligencia de negocios. La combinación de estrategias de minería de datos, corrección y análisis lingüístico consigue resultados óptimos en el campo de la gestión de la información y la subsiguiente toma de decisiones.

7.1.4. Análisis DAFO

Oportunidades:

- **Exportar experiencias del mercado nacional** a otros mercados (TA integrada en procesos productivos, ...)
- El **mercado latinoamericano puede ser una gran oportunidad para proyectos** entre español y portugués (brasileño). Igualmente existe una gran oportunidad en el auge que está teniendo el español en EE.UU.
- El gran auge que están teniendo las **redes sociales**, juntamente con el procesamiento del Big Data sitúan las industrias de la lengua en una excelente posición para, por un lado, explorar nuevos ámbitos de actuación (análisis del sentimiento) y por otro, avanzar en la mejora de los sistemas a través de la obtención de recursos básicos (datos) para avanzar en la mejora de los sistemas (traducción automática).
- El **gran número de agentes implicados en el sector** (centros de investigación, centros de desarrollo de recursos básicos, empresas desarrolladoras de productos, empresas de integración, empresas de servicios) debe ser visto como una gran oportunidad. Existe una gran oportunidad en integrar las TIL y sus agentes en clusters o polos empresariales en los cuales pueden aportar un gran valor añadido: sector editorial, medios de comunicación, sector ocio, sector turismo, cámaras de comercio, comercio exterior, etc.



Fortalezas:

- **Número de agentes** (entre empresas y grupos de investigación de universidades) importante. Existe una gran experiencia en el ámbito de las TIL. Se trata de un sector muy desarrollado, maduro y competente.
- **Administraciones públicas** como grandes consumidoras y dinamizadoras de las TIL.
- La **oferta propia** de soluciones es muy significativa, se cuenta con muchos productos nacionales y con importantes experiencias que podrían ser exportadas a nuestros ámbitos más próximos (Latinoamérica, Portugal, Europa).
- El sector español cuenta con **soluciones y especialización** en lenguas "no oficiales" por el momento en la CE.

Debilidades:

- **Dimensión de las empresas:** PYMES o microempresas. En algunos casos empresas de 1 o 2 personas.
- **Visibilidad** del sector en general, de su potencial y de su transversalidad.
- **Poca diferenciación** entre diversas tecnologías, servicios o productos.
- Estrategia de **comercialización poco definida**, existen grandes dificultades para llegar al mercado.
- Falta de **líneas de financiación** que ayude a la internacionalización de las empresas y acometer proyectos de I+D+i
- Falta de **interoperabilidad y compatibilidad** entre los diversos agentes y sus tecnologías.
- **Poca inversión** en investigación
- **Dificultades para crecer** y para internacionalizarse por parte de las empresas.
- La mayoría de las empresas han indicado explícitamente la necesidad de establecer líneas de apoyo para el sector.

Amenazas:

- Potencia de las **grandes empresas extranjeras** del sector: Google, Asia OL, Microsoft, IBM, etc.
- Las grandes empresas extranjeras pueden entrar en el **mercado local español** con riesgo de trivializar el tratamiento lingüístico de nuestras lenguas.



La tecnología es válida para cualquier sector, es decir, no están orientadas a mercados concretos. Este hecho puede considerarse como una debilidad si denota la falta de concreción de una estrategia de implantación y personalización de las TIL en entornos específicos, estrategia que debería incluir la intervención activa de diversos agentes con sus recursos básicos correspondientes.

7.1.5. Recomendaciones

- Aumentar la **visibilidad del sector**. Crear un repositorio donde fácilmente puedan encontrarse referencias a los diversos agentes, casos de éxito, etc.
- Facilitar la **interconexión e interoperabilidad entre los diversos agentes/tecnologías**. Dinamizar clusters y fomentar la transversalidad de las TIL en diversos polos económicos (medios de comunicación, administraciones públicas, contenidos, sector editorial, turismo, educación, etc.)
- Facilitar que las administraciones públicas **adquieran servicios y tecnologías de forma ‘adecuada’**. Es decir, que se permita la confluencia de diversos actores en las adquisiciones y que se prime siempre (por delante del criterio estrictamente económico) la posibilidad de diseñar proyectos de implantación tecnológicos donde se puedan visualizar valores añadidos para los diversos actores: generación de recursos, curvas de aprendizajes, retroalimentación de sistemas, etc.
- Acceso público a recursos textuales para facilitar el desarrollo de nuevos proyectos (Open Data).
- Habilitar **líneas de financiación para PYMES** para abordar programas de internacionalización y también programas de I+D+i.
- Estimular la **relación universidad** (centros de investigación) y **empresa**.
- Impulsar el **crecimiento de las empresas**.
- Impulsar acciones encaminadas a que las universidades integren en sus planes de estudios la formación de los **nuevos perfiles profesionales** que se requieren para el desarrollo de las TIL (posteditores, procesadores del lenguaje, tratamiento de datos, etc.)

7.2. Transferencia de conocimiento sector investigador-industria

El informe “Language Technologies LT2013: Status and Potential of the European Language Technology Markets” impulsado por el Foro de Industrias de las Tecnologías de la Lengua LT-Innovate **explora diferentes tipos de modelos de colaboración y los escenarios que ilustran su uso potencial** indicando aquellos que pueden ser adecuados para la financiación pública. Así se distingue, en primer lugar, como escenario ideal aquél en el que se entreteje una infraestructura núcleo de filosofía abierta y diseñada para una red de organizaciones que participan en diferentes puntos de la cadena de valor. Este es el escenario ideal para una transferencia sana y natural entre la academia y un ejemplo de este escenario es el **proyecto KNIME** (www.knime.org), un ecosistema creado sobre una plataforma open-source para el dominio científico.



A continuación, el mencionado informe propone promover como **escenarios “posiblemente adecuados”** los siguientes:

- **Alianzas abiertas con compromiso a corto plazo**, entre partners que comparten capital intelectual, riesgos en i+D y beneficios. Adecuado para proyectos de desarrollo más especulativos y que se encuentran en estado inicial de desarrollo. Este tipo de alianzas son interesantes para proyectos a pequeña escala y no tienen por qué utilizar tecnología de código abierto. Un ejemplo de este escenario es la alianza AppCampus (<https://www.appcampus.fi/>), alianza entre las más importantes compañías de Information Technologies y *Start-ups* en Finlandia.
- **Ecosistemas de compañías con incidencia similar en el mercado** en los que se **intercambia tecnología entre componentes** de una red con áreas de aplicaciones que se solapan. Se trata de ecosistemas guiados por las necesidades industriales. En este tipo de iniciativas la interoperabilidad y la estandarización ocuparán un lugar importante; es probable que no den lugar a tecnología abierta. Ejemplo de este escenario es el ecosistema SpeechCloud (www.ispeech.org) conjunto de empresas que desarrollan y distribuyen ASR y TTS sobre una plataforma en la nube con el objetivo de competir con la empresa Nuance.
- **Alianzas entre compañías con incidencia similar en el mercado que comparten y desarrollan tecnologías complementarias**. Se trata de un tipo de alianza tradicional en el que ninguna de las compañías prevalece sobre las demás y que pueden afrontar riesgos a largo plazo juntos. Ejemplo de este tipo de alianzas es la establecida entre Nuance e IBM (Nuance & IBM. <http://www-03.ibm.com/press/us/en/pressrelease/33726.wss>) para integrar tecnologías inteligentes para la búsqueda de soluciones en el área de la salud.
- **Ecosistemas asimétricos en los que una compañía fuerte en el mercado trabaja con una red de organizaciones** que participan en diferentes puntos de la cadena de valor. En este tipo de ecosistemas, las compañías pequeñas se benefician en el sentido de que tienen acceso a mercados más amplios, mayor número de clientes sin perder autonomía. Para las compañías más fuertes, supone ampliar sus capacidades sin ningún riesgo. Ejemplo de este ecosistema es Salesforce Marketing Cloud Ecosystem⁶⁸.



El informe añade como alianza no deseable aquella que se establece entre una compañía fuerte y una compañía menor que desarrolla tecnologías innovadoras.

⁶⁸ www.microsoft.com/en-au/dynamics/crm.aspx?WT.mc_id=DynGB_en_au_SEM_GOOG&DYNCRM-SEARCH&CR_CC=200578851&WT.srch=1

7.2.1. Transferencia en los organismos públicos de investigación

Según el Informe de la Encuesta de Investigación y Transferencia 2012 y 2013 de las Universidades Españolas⁶⁹ llevado a cabo por el Grupo de Trabajo de Indicadores, RedOTRI y RedUGI en 2015, **las formas de transferencia de conocimiento son las licencias sobre resultados, la creación de empresas de base tecnológica (EBT o *spin-off* académicos) y la interacción con terceros en actividades de I+D y apoyo técnico**. La forma de transferencia de conocimiento más habitual en las universidades españolas es la interacción con terceros, fórmula con la que la industria puede acceder a la capacidad instalada en la universidad y obtener derechos de explotación sobre conocimiento nuevo relevante para los procesos de innovación social y económica.

La protección de los resultados de la investigación está mayoritariamente **a cargo de las Oficinas de Transferencia de Resultados de Investigación (OTRIs)**. El mecanismo de protección habitual es la patente, pero en el caso de resultados no patentables, como el software, existe otra vía de protección. En España esta vía es el derecho de autor, según regula la legislación de propiedad intelectual. La titularidad de la universidad sobre los resultados de investigación ocurre en aquellas actividades que son **objeto de subvención pública o mecenazgo privado**. En el caso de los contratos con empresas, la titularidad se negocia entre las partes (según datos sobre encuestas, solo un 12% recae sobre la universidad).

Las ***spin-off* académicas** son empresas generadas para explotar los resultados de la investigación universitaria. En principio, constituyen no sólo un mecanismo muy efectivo para transferir resultados a la sociedad, sino también una fórmula muy atractiva para estas instituciones para contribuir a la renovación del tejido productivo local con actividades de proyección global y de alto valor añadido, además de retener talento y capital intelectual, frecuentemente procedente de las mismas universidades. Las *spin-off* son también el mecanismo de transferencia más complicado puesto que deben atender tanto a los aspectos ligados al conocimiento a transferir y al mismo tiempo tener en cuenta todas las cuestiones del funcionamiento empresarial, que además, debe operar en un entorno global, aun siendo de pequeño tamaño. **Muchas universidades cuentan con unidades con ese cometido** de apoyo empresarial que son diferentes de las OTRI. Las *spin-off* académicas suelen ser empresas con cierta fragilidad, como cualquier negocio de nueva creación, por lo cual es importante conocer el grado de supervivencia de éstas al cabo de los años, aunque a veces pueden estar en situación latente o con bajo nivel de actividad.

Las **ampliaciones de capital** también son una señal de validación de las *spin-off* puesto que durante los primeros años de vida requieren validar un producto, modelo de negocio y mercado, todo con un alto grado de innovación y con una necesidad de un fuerte ritmo de crecimiento. En cuanto al personal investigador que ha promovido empresas, éste ha sido de 132 personas en 2012 (cifra menor



En 2012 el número de *spin-off* que ampliaron capital fue de 33 y en 2013 de 45. La falta de capital semilla y de capacidad de gestión puede ser el origen de la falta de evolución de las *spin-off*, de hecho, la inversión privada en *spin-off* en 2012 y 2013 fue solo de 10 millones de euros.

⁶⁹ <http://www.crue.org/Paginas/Encuesta-de-Investigacion-y-Transferencia.aspx>

que en 2011, probablemente motivado por la mayor dureza de la crisis económica en España), y en 2013 de 198 investigadores⁷⁰.

Según se deriva de los datos globales de párrafos anteriores, las *spin-off* centradas en el PLN o la TA son **pocas en términos absolutos**, y surgen también en la mayoría de casos de proyectos o grupos de investigación:

- Verbio - UPC
- Cilenis - Universidad de Santiago
- Semantia Lab - UNED
- Eleka Ingeniaritza Linguistikoa SL - UPV/EHU
- SciLing - UPV
- Prompsit - Universidad de Alicante

Hay **dos casos de empresas nacidas como *spin-off* universitarias** que han conseguido consolidarse. El primer caso es Daedalus, fundada en 1988 por profesores de la UPM y la UAM, que es hoy una empresa privada con clientes de varios sectores: Telefónica, PRISA, Hispasat, Unidad Editorial, Iberdrola, Vocento, Amper, RTVE, Eutelsat, Hibü, Instituto Cervantes o Digimind. El segundo caso es Indisys, fundado por investigadores de la Universidad de Sevilla, y que ha sido recientemente adquirida por Intel por 20 millones de euros para su división de diálogo y asistente virtuales⁷¹.

7.2.2. Transferencia de conocimiento a la industria

Las relaciones entre el entorno empresarial (empresa) y el entorno científico (organismos públicos de investigación), presentan **dificultades intrínsecas derivadas de los diferentes objetivos, estructura y medios de cada una de ellas**, por lo que no es extraño que se hayan ideado diversas formas de mejorar su situación.

Las barreras a la cooperación pueden **visualizarse gráficamente** por la dificultad en adaptar todos los objetivos planteados:

- El **objetivo de investigación básica de un centro público** difícilmente es aceptado como tal por una empresa, y lo mismo sucede en el caso de la industrialización de un determinado producto necesario para la empresa, pero alejado de los intereses de un centro público de investigación.
- En otros casos, la brecha existente puede paliarse por un **mejor conocimiento de los objetivos y las limitaciones de cada uno**.



⁷⁰ <http://www.crue.org/Paginas/Encuesta-de-Investigacion-y-Transferencia.aspx>

⁷¹ http://cincodias.com/cincodias/2013/09/13/tecnologia/1379096387_633326.html

Un estudio encargado por la Comisión Europea (Dankbaar, 1993), concluyó que el proceso de **transferencia de tecnología depende de los siguientes factores**:

- **Base industrial en que se desarrolla**, entendida ésta como el nivel de desarrollo, sofisticación tecnológica, naturaleza y actividad de la empresa.
- **Tamaño de la empresa**, pues las Pymes se caracterizan por la dificultad de acceder directamente a los recursos tecnológicos de la universidad, siendo más efectivo realizar el enlace a través de organismos intermedios. Por su parte, las grandes empresas desarrollan mecanismos de enlace directos con las instituciones del ámbito científico debido a su mayor capacidad investigadora.
- **Identificación del sector de actividad**, que define las tecnologías que son más relevantes para la empresa, la estructura del sector y el nivel de competencia existente.
- **La región de ubicación de la empresa**, que refleja el nivel de desarrollo económico, la calidad de los conocimientos básicos y las restricciones de carácter administrativo del entorno.

Según este estudio, los **principales factores que el entorno científico debe desarrollar para asegurar el éxito** de la transferencia de tecnología al ámbito empresarial pueden agruparse en:

- Ampliar la oferta tecnológica.
- Desarrollar acciones de marketing.
- Gestionar de forma eficiente la transferencia de tecnología.
- Utilizar instrumentos de apoyo financiero.

Generar empresas de base tecnológica (spin-offs).

Los grupos de investigación españoles tienen **cierta actividad de colaboración directa con empresas españolas**. Los datos suministrados (ver Anexo, tabla 5) muestran contratos de servicios y suministro de datos así como, en menor medida, de colaboración para el avance tecnológico en la mejora de componentes desarrollados por las empresas. Los datos suministrados por la universidad en los últimos años han servido para el desarrollo de: Correctores ortográficos; Herramientas terminológicas y Procesadores lingüísticos;

En general, la colaboración entre ambos sectores se da en el **marco de proyectos con financiación externa**, tanto de convocatorias de ámbito nacional como, en su mayoría europeas. A continuación, se presenta una muestra representativa:

- Synapse Développement (Francia) Proyecto READERS;
- Higher Functions, Lda (Portugal) Proyecto QT Leap;
- LexisNexis (Holanda) Proyecto NewsReader;

Un ejemplo relevante lo proporciona el grupo IXA de la Facultad de Informática de San Sebastián (Universidad del País Vasco (EHU/UPV), que ha detallado la siguiente colaboración con empresas en el marco de proyectos europeos.



- ScraperWiki (Reino Unido) Proyecto NewsReader;
- Synerscope (Holanda) Proyecto NewsReader;
- Synthema (Italia) Proyecto OpeNER y ProyectoKyoto;
- Olery (Holanda) Proyecto OpeNER;
- Irion Technologies (Holanda) Proyecto Kyoto;
- European Centre for Nature Conservation (Holanda) Kyoto;
- Synapse Développement (Francia) Proyecto READERS;
- Higher Functions, Lda (Portugal) Proyecto QT Leap;
- LexisNexis (Holanda) Proyecto NewsReader;
- ScraperWiki (Reino Unido) Proyecto NewsReader;
- Synerscope (Holanda) Proyecto NewsReader;
- Synthema (Italia) Proyecto OpeNER y Proyecto, Kyoto;
- Olery (Holanda) Proyecto OpeNER;
- Irion Technologies (Holanda) Proyecto Kyoto;
- European Centre for Nature Conservation (Holanda) Kyoto;



7.2.3. Análisis DAFO

A partir de la información y comentarios enviados por algunos de los grupos de investigación en PLN/TA de las universidades españolas, se ha confeccionado el siguiente DAFO:

Oportunidades

- **Oportunidades de trasvase** de tecnología y know how.

Fortalezas

- **Oferta de herramientas propias** desarrolladas por los grupos de investigación y utilizadas por las empresas.
- **Equipos humanos**, de centros y grupos de investigación, vinculados a la investigación, consolidados.
- **Colaboraciones con centros de investigación** a través de proyectos europeos y colaboraciones entre centros y universidades del ámbito autonómico, estatal e internacional.

Debilidades

- **Empresarial** (demanda potencial) de las tecnologías lingüísticas (TA y PLN), existentes en el entorno investigador.

- La investigación resulta estar **más avanzada que la necesidad** o demanda del mercado.
- **Inversión elevada** –coste y tiempo- necesaria para el desarrollo de nuevas herramientas tecnológicas.
- La forma de transferencia de conocimiento más habitual, también en general, es la **interacción con terceros en actividades de I+D y apoyo técnico**. Con esta fórmula la industria accede a la capacidad instalada en la universidad, pero en la actualidad no hay infraestructuras identificable de investigación de PLN y TA de las que la industria se pueda beneficiar.
- Hay **muy pocas spin-off en la universidad española**⁷², 633 en 2013, de las cuales unas 6 estarían dedicadas a PLN y TA, probablemente debido a la complejidad de esta fórmula de transferencia, aunque algunas spin-off se han desarrollado como empresas (destacables Daedalus e Indisys) y una de ellas ha sido adquirida por Intel por 20 millones de euros (Indisys).

Amenazas

- **Reducción de las ayudas** a la investigación por parte de la Administración Pública.
- **Alta competencia** de herramientas desarrolladas en el extranjero



7.2.4. Formación de profesionales de PLN y TA en España

Según los libros blancos editados por META-NET, la formación básica en lingüística computacional se realiza en España en el marco de las licenciaturas en Lingüística, Filología o Traducción, que incluyen esta materia como troncal, o en alguna ingeniería informática. **Los cursos de postgrado ofrecen una formación profesional más específica.** Hay varios programas de doctorado que ofrecen módulos o asignaturas relacionadas con el lenguaje o el procesamiento del habla, según la SEPLN, hay 18 programas de doctorado en España, la mayoría en escuelas de informática, donde se da formación en PLN. Los masters a menudo se ofrecen de forma interuniversitaria, por grupo de universidades, ya sea a nivel estatal o a nivel europeo. También se ofrecen módulos de Tecnología Lingüística a los estudiantes de otros másteres o cursos de doctorado, sobre todo en traducción (p. ej. en la Universidad Autónoma de Barcelona, Universidad de Alicante, Universidad de Castellón, Universidad de Valencia y Universidad de Granada). Algunas universidades como la Universidad Politécnica de Cataluña participan en el Máster Europeo en Lengua y Habla patrocinado por ELSNET (Red Europea de Excelencia en tecnologías del lenguaje humano).

La Universidad del País Vasco participa en un **consorcio de universidades que ofrecen un máster europeo Erasmus Mundus** en tecnologías del lenguaje y la comunicación y también un máster propio, el Máster Universitario en Análisis y

“El uso de las herramientas de traducción y traducción automática crecerá gracias al volumen de material disponible para traducir. El trabajo de traductor será más diverso con un incremento del uso de las herramientas y máquinas de traducción.” GALA (Globalization and Localization Association). 2013

⁷² <http://www.crue.org/Paginas/Encuesta-de-Investigacion-y-Transferencia.aspx>

Procesamiento del Lenguaje. La Universidad Autónoma de Barcelona ofrece otro Máster Internacional Erasmus Mundus en Procesamiento del Lenguaje Natural y Tecnologías del Lenguaje Humano, en colaboración con universidades extranjeras. En Galicia existe el Máster Interuniversitario en Lingüística Aplicada que, desde hace 3 años, imparten conjuntamente las universidades de Vigo, A Coruña y Santiago de Compostela, con asignaturas obligatorias de tecnologías lingüísticas y un módulo específico sobre este tema.

En Latinoamérica la situación formativa no es comparable con la española.

En Brasil no hay cursos universitarios sobre PLN o lingüística computacional. Sin embargo, sí hay algunos cursos de doctorado en al menos cinco universidades donde se da formación en PLN. El área ha experimentado una internacionalización ya que los estudiantes de doctorado pueden obtener financiación para completar total o parcialmente sus estudios de doctorado en el extranjero, pero se desconoce si existe compromiso de ejercer durante un período de varios años en Brasil.

Otro recurso formativo interesante son **cursos en línea abiertos y masivos (massive open online course o MOOCs)**. Estos cursos ofrecen una oferta formativa abierta a cualquier estudiante y en la mayoría de casos es posible obtener una certificación final. A nivel internacional existen dos MOOCs sobre PLN muy conocidos, alojados en Coursera:

- *Natural Language Processing* de la Universidad de Columbia, impartido por el Prof. Michael Collins (80-100 h. de trabajo)
- *Natural Language Processing* de la Universidad de Stanford, impartido por los Profs. Dan Jurafsky y Christopher Manning (60-80 h. de trabajo)

En España existe el **MOOC Nuevos retos en las Tecnologías del Lenguaje Humano**, alojado en la Universidad de Alicante, en el que participan 19 profesores.

La **lista de cursos formación y masters** vinculados al procesamiento de lenguaje natural en España más completa se puede encontrar en: http://www.sepln.org/category/investigacion/grupos_investigacion/ (también en la Tabla número 6 “Formación en NLP en España” del Anexo).

En cuanto a las becas para PLN, no existe ninguna oferta específica.

- Respecto a becas predoctorales existen a nivel nacional las becas de Formación de Personal Investigador o de Profesorado Universitario (FPI o FPU), en el caso de Cataluña las becas de Formación de Investigadores (FI).
- Respecto a becas postdoctorales, existen algunos investigadores en PLN en los programas Ramón y Cajal, Juan de la Cierva y Torres Quevedo y en cuanto a becas de máster, se obtienen mediante convocatorias de Becas de carácter general y de movilidad para estudios universitarios, Becas de colaboración y Préstamos para realizar estudios de posgrado de

A nivel internacional, son destacables los casos de Reino Unido y Alemania que cuentan con estudios de grado y posgrado en PLN en casi todas las universidades.



Máster universitario o de Doctorado⁷³, todas ellas financiadas por el Ministerio de Educación, Cultura y Deporte, además de las becas generalistas ofrecidas por las Comunidades Autónomas.

Según el último “Informe de la encuesta de investigación y transferencia 2012 y 2013 de las universidades españolas”, la transferencia de conocimiento (TC) vinculada a la actividad de I+D+i con empresas ha presentado en 2012 un resultado similar al del año anterior (557M€), **pero en 2013 ha vuelto a retroceder ligeramente hasta los 533M€**. La tendencia a la baja de este indicador se mantiene desde el inicio de la crisis económica pero no tiene la profundidad que ha tenido la financiación competitiva. En 2013 se aprecia una recuperación de la componente de investigación por encargo, aunque **no es suficiente para contrarrestar la disminución en los contratos de apoyo técnico** y, sobre todo, en la contracción de la financiación pública a la investigación colaborativa con empresas.

Aunque, de una forma muy leve, la interacción con empresas **se va abriendo cada vez más a mercados internacionales**, tanto en la Unión Europea como en el resto del mundo, representando ya un 15% del total de la contratación. La propia región sigue siendo el principal mercado de las universidades, con algo más de la mitad de la contratación. Por su parte, la tecnología ligada a explotación de resultados, aunque mejora en cuanto al número de contratos de licencia, ello **no se está traduciendo en una mejora de los ingresos por acuerdos de propiedad intelectual/industrial**, que sólo han sumado 2,4 M€ en 2012 y 2,1 M€ en 2013. Quizá sea debido a que las licencias se están orientando a empresas spin-off y a PYMES europeas, con plazos más largos para conseguir retornos por explotación. Este tipo de mecanismo de transferencia sigue siendo residual en España, y, además, se concentra en sólo la mitad de las universidades⁷⁴.

El crecimiento en solicitudes de patentes experimentado en los últimos años no se está traduciendo en un aumento apreciable de licencias ni de sus retornos económicos.

7.2.5. Análisis DAFO

Oportunidades

- La **gran diversidad de aplicaciones** PLN ofrece una gran oportunidad de avanzar en numerosos campos industriales/empresariales (automoción, TICs)
- Introducir los aspectos de **Traducción Automática** (a nivel tecnológico), en los estudios de universitarios de Traducción e Interpretación.

⁷³ <https://sede.educacion.gob.es/catalogo-tramites/becas-ayudas-subvenciones/para-estudiar/master.html>

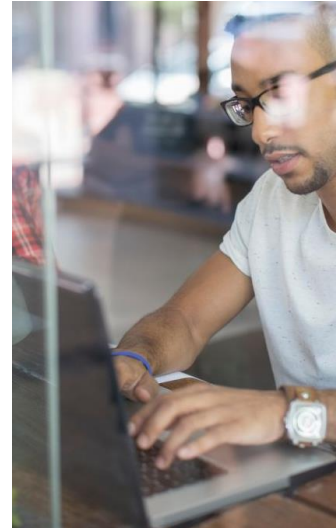
⁷⁴ Informe de la encuesta de investigación y transferencia 2012 y 2013 de las universidades españolas
http://www.crue.org/Documents/Informe%20Encuesta%20ITC%20RedOTRI_2012_y2013.pdf

Fortalezas

- Tanto el PLN como la TA están en posición de **liderar una nueva dimensión** en las aplicaciones informáticas del futuro (comunicación hombre-maquina).

Debilidades

- A pesar del gran desarrollo tecnológico y oferta formativa en PLN, a menudo resulta ser **desconocido para los profesionales**.
- **No existen oportunidades educativas**, vinculadas a postgrados y masters, en todas las universidades donde se imparten estudios de PLN y TA
- Hay al menos **18 programas de doctorado** que ofrecen alguna asignatura o módulo de PLN y algunos máster (~4) específicos, algunos internacionales (Erasmus Mundus y ELSNET), pero no hay estudios sobre estas áreas claramente identificados, mientras que Reino Unido o Alemania cuentan con estudios de grado y posgrado en PLN en muchas universidades. En España, la formación de grado se realiza en el marco de los estudios de lingüística, filología, traducción e informática.
- **No hay oferta de becas específicas para PLN o TA** y los doctorados se ven forzados a una diáspora, algunos de ellos trabajando en Watson (IBM), Nuance o Google.



Amenazas

- **Gran oferta educativa vinculada a la TA y PLN en Europa y EEUU.**

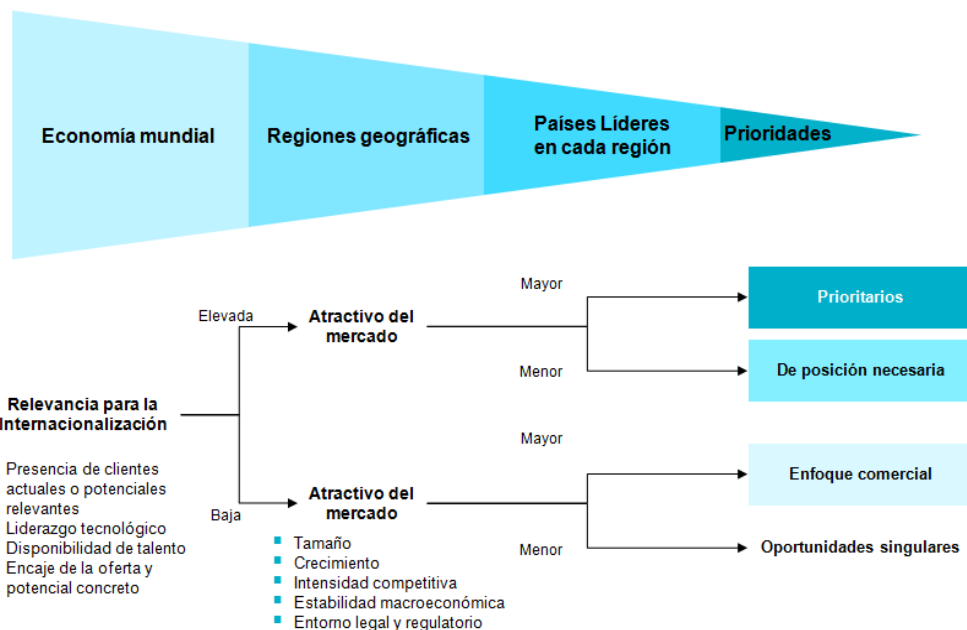
7.3. Internacionalización

7.3.1. Concepto de internacionalización

El hecho de construir un software que pueda responder a convenciones culturales distintas, aspecto éste cada día más importante en la denominada “aldea global de McLuhan”, se conoce como internacionalización. Multinacionales españolas como pueden ser BBVA, Santander, Telefónica han hecho de “**tractores para la internacionalización**”. Grandes consultoras, como Indra, con su expansión internacional está realizando labores de internacionalización con las tecnologías de la información y por lo tanto podría llegar a ser “tractora” en algunos aspectos relacionados con este tipo de tecnologías.

Disponer de **soluciones diferenciales** y de **alto valor añadido** que no existen en la geografía es un mecanismo de entrada e implantación local. Con la siguiente imagen se plasma la expansión de un **proceso de internacionalización** para una tecnología de este tipo.

Internacionalización:
proceso estratégico por el cual se desarrollan capacidades para hacer negocios en países fuera de nuestro entorno natural.



7.3.2. Líneas de acción

Revisando la información disponible y analizando las razones de internacionalización, en este epígrafe se pueden enmarcar **dos caminos independientes en cuanto a los desarrollos que se puede acometer de cara a favorecer la expansión en otros países.**

7.3.2.1. Herramientas del mercado: Creación de una nueva herramienta

La creación de nuevas herramientas, permiten por un lado el **desarrollo de utilidades que puedan facilitar la vida** en tanto en cuanto se pueden usar, sacando partido al interactuar con la máquina en la búsqueda de:

- Palabras en contexto.
- Corrector ortográfico para una palabra.
- Corrector ortográfico para un fichero.
- Diccionario de sinónimos.



Ejemplo de empresas del sector con herramientas en el mercado:

- Creative Virtual (UK):** empresa que desarrolla asistentes de clientes virtuales (VCA) y que en el contexto de una conversación cliente –VCA, la máquina es capaz de transmitir estados emocionales.
- Linguamatics (UK):** empresa que desarrolla herramientas con algoritmos lingüísticos de minería de textos para la extracción de información.

- **Acrolinx (DE):** empresa que desarrolla herramientas de ayuda a la creación de contenidos con tecnología de PLN.
- **Eptica-Lingway (FR):** empresa especializada en la extracción de información de documentación como currículums para dar apoyo a tareas de recursos humanos.
- **SparkCognition (EEUU):** empresa dedicada al reconocimiento de patrones y tendencias, y el aprendizaje automático para:
 - Detectar amenazas en tiempo real de ataques potenciales en internet y servicios asociados (por ejemplo, pago telemático).
 - Fallos de rendimiento de sistemas informáticos.
- **SpeakToit (api.ai) (EEUU):** compañía que desarrolla interfaces de comunicación utilizando el procesamiento del lenguaje natural (texto y voz) y el aprendizaje automático, para ser integrados otros sistemas ya existentes.
- **Amy (X.ai) (EEUU):** empresa que desarrolla y comercializa asistentes de programación virtual que organizan y agendan reuniones de trabajo tras la lectura (en lenguaje natural) de los correos electrónicos de los usuarios a reunir.
- **SDL (EEUU). Herramienta (Trados Studio 2015);** un entorno completo para profesionales del lenguaje para editar y revisar traducciones, gestionar proyectos de traducción, organizar terminología corporativa y conectarse a motores de traducción automática.
- **Watson Analytics (IBM, EEUU) es una herramienta de análisis de datos avanzada (evolución de Cognos Insight)** con las siguientes características:
 - Se ejecuta en la nube: primera herramienta de IBM que presenta esta funcionalidad.
 - Aporta a los usuarios de la misma consulta y respuestas en lenguaje natural.
 - Presenta una continua evolución, con entregas de versiones periódicas.
 - Está formada por 4 módulos:
 - o **Explorador:** Módulo que permite cargar archivos planos (.csv, xls., xlsx) y realizar sobre ellos consultas en lenguaje natural a través de un cuadro de texto y modificar palabras clave para modificar la consulta. También accede a redes sociales (twitter) para realizar consultas sobre la información almacenada.
 - o **Predicción:** Módulo que permite realizar predicciones de los datos base de forma gráfica y genera patrones de comportamiento. Para que sea eficaz, el usuario de este



módulo debe tener un conocimiento en el análisis de datos y estadística. No actúa como una caja negra, si no que permite cierta flexibilidad al usuario para realizar sus predicciones y establecer patrones.

- o **Ensamblador:** Módulo que realiza el montaje de los patrones y predicciones elaboradas en un archivo (infografía) para que sea compartido y analizado por otros usuarios analíticos. Es el módulo más limitado en su funcionalidad: los gráficos establecidos por los módulos de exploración y predicción solo pueden ser copiados en éste para su visualización.
- o **Refinador:** Módulo que se ha añadido recientemente a la herramienta para crear grupos y cálculos básicos.

7.3.2.2. Herramientas en el mercado: programas de análisis en redes sociales

Hay que resaltar **el éxito en herramientas orientadas al procesamiento y análisis de los flujos masivos de información encontrada en las redes sociales**. El manejo de toda esta información no solo requiere de conocimiento específico de las herramientas con las que se vaya a trabajar, además implica análisis lingüístico con herramientas de PLN, tales como:

- **Análisis morfológico:** determinar la forma, clase o categoría gramatical de cada palabra de una oración.
- **Análisis sintáctico:** relaciones de concordancia y jerarquía que guardan las palabras agrupándose entre sí en oraciones simples y compuestas.
- **Análisis semántico:** estudio de las palabras en función de su significado.
- **Análisis pragmático:** estudio de las palabras en función del contexto extra-lingüístico.

El **auge de las redes sociales** deja una huella digital que recoge nuestros usos, costumbres, intereses, opiniones, gustos, etc.; con el pequeño inconveniente de que analizarlos se convierte en una tarea titánica si no se dispone de las [herramientas](#) adecuadas. Esta necesidad ha dado un nuevo impulso a la «ciencia de los datos».

Extraer significado de la conversación de manera automática requiere herramientas con PLN, como Lynguo, Watson, etc..., que permiten categorizar, analizar el sentimiento, las emociones o incluso la concienciación.

7.3.2.3. Modelos de internacionalización

A la hora de efectuar la internacionalización de las empresas que crean este tipo de herramientas, que las implantan, desarrollan o analizan, se debe tener en cuenta los modelos de internacionalización a seguir por la empresa; llevando a cabo alguna de las siguientes **acciones**:

- Iniciar la exportación de los productos seleccionados.

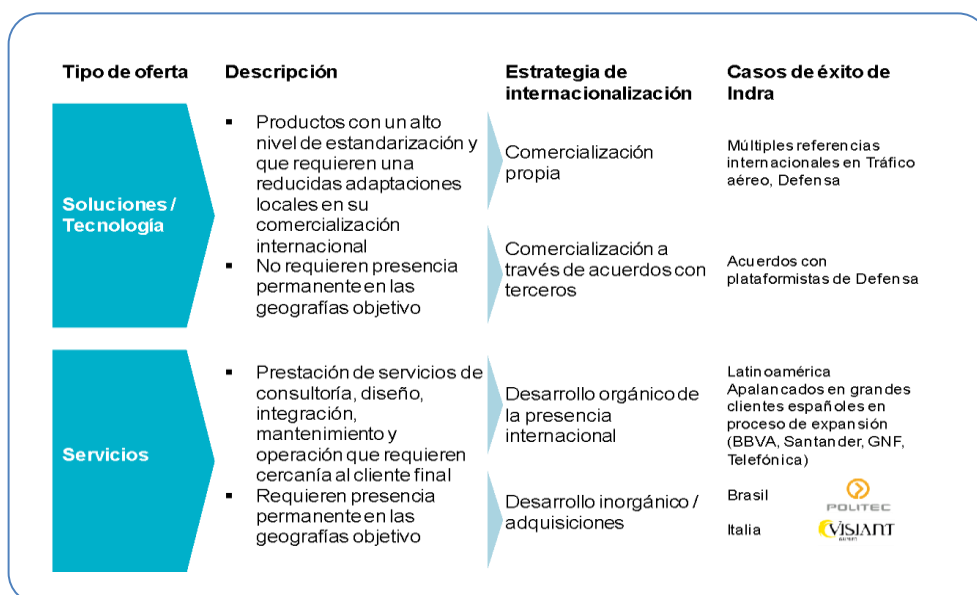


- Identificación de la industria a la que se pretende llegar.
- Consolidar mercados exteriores.
- Buscar nuevos mercados emergentes o nichos de mercado por descubrir.
- Localizar nuevos proveedores, con los que ir de la “mano”.
- Localizar subcontratistas “locales”.
- Buscar nuevos productos, incorporando las mejoras investigadas.
- Buscar nuevas aplicaciones a implantar o a desplegar.
- Desarrollar contactos exteriores para lograr la penetración adecuada en los mercados que interesen.
- Efectuar prospecciones y estudio de oportunidades reales de negocio con los instrumentos que se detallarán más adelante.
- Búsqueda socios tecnológicos nacionales, internacionales, locales.
- Búsqueda socios comerciales que conozcan perfectamente el mercado.
- Búsqueda socios capital, para lograr la financiación parcial o total de las acciones a realizar.
- Desarrollar la cooperación internacional con el intercambio de conocimiento – experiencia.
- Acuerdos de colaboración con agentes en los países.
- Fusiones-absorciones con empresas que desarrollan este tipo de tecnología.



7.3.2.4. Estrategias de Internacionalización

Las estrategias de internacionalización conocidas y llevadas a cabo por Indra, en casos de éxito, se muestran en la **siguiente figura**:



En esta imagen se refleja la experiencia de Indra como empresa referente en las tecnologías de la información.

7.3.2.5. Procesos de internacionalización

En los procesos de internacionalización existen **obstáculos que las empresas pueden encontrar** y que indican la parte negativa de la internacionalización de las empresas. Seguidamente se enumeran los principales escollos detectados estudiando la información disponible de las empresas del sector:

La propia empresa

La internacionalización no es un proceso fácil y el gran problema no es poder hacerlo, es querer hacerlo y hoy por hoy con la globalización, los mercados maduros, con la gran competencia que existe, necesitan de **equipos profesionales que lleven a buen puerto una de las estrategias más rentables a largo plazo**.

Diferencias culturales

Es preciso tener en cuenta el país al que van destinadas las mercancías tanto desde punto de vista del idioma como de la cultura, ya que su forma de pensar y de valorar los negocios puede diferir de manera importante. Las **diferencias culturales pueden afectar a los gustos de los consumidores**, a la forma de hacer negocios o a la comunicación con el mercado.

Problemas de localización

Las distancias suponen grandes problemas tanto por el tiempo necesario para la llegada de mercancías a su destino como por el encarecimiento del transporte que repercute en el coste del producto. Tampoco hay que olvidar que las **grandes distancias dificultan la comunicación directa** de la empresa con sus distribuidores y sus consumidores en el mercado exterior.

Situación socioeconómica del país receptor

La situación económica del país puede **afectar a la capacidad de compra de los productos** haciendo más difícil su comercialización. Además la situación económica, política y social puede generar inestabilidad y situaciones de riesgo muy comunes en el comercio exterior, es lo que se conoce como riesgo país, que si es posible conviene evitar.

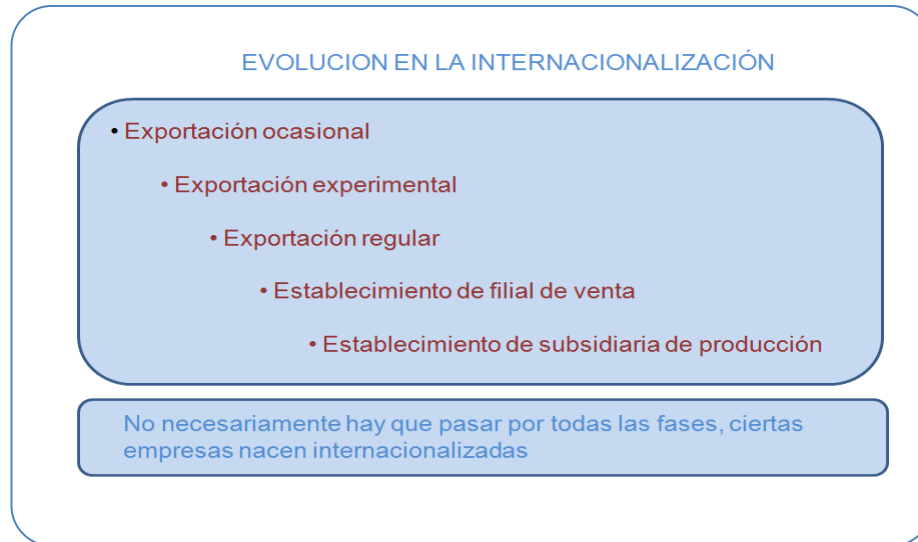
Barreras arancelarias

Los gobiernos, con objeto de proteger su industria **nacional gravan los productos importados con derechos arancelarios**, los cuales suponen un incremento en los precios de los productos importados haciéndolos menos competitivos. En otras ocasiones se utilizan otros tipos de instrumentos proteccionistas de similares efectos: contingentes, cupos, cuotas a la importación.



Es labor de la Dirección de estas empresas buscar ayuda de Instituciones, Asociaciones, buenos asesores externos que guíen el proceso y formen al líder o al equipo que lidere el proceso

Constantemente, las empresas deben desarrollarse y buscar soluciones a los muchos problemas que se irán encontrando a lo largo del tiempo. Si una empresa no evoluciona, desaparece. En la siguiente figura se plasma la evolución natural de los procesos de internacionalización:



Evolución de la Internacionalización

7.3.3. Instrumentos

A continuación, se presenta una muestra de instrumentos para llevar a cabo los procesos de internacionalización en las empresas con este tipo de tecnología:

7.3.3.1. ICEX (Instituto Español de Comercio Exterior)

El ICEX ⁷⁵ fomenta una política activa de colaboración con otras instituciones cuya finalidad es apoyar la internacionalización de la empresa, con el objetivo de establecer **programas específicos de actuación conjunta**, dentro de una política de promoción común, y obtener un mejor impacto en la asignación de recursos.

El ICEX, mantiene una colaboración bilateral con todas las asociaciones sectoriales, cámaras de comercio, etc.

Para ello, el ICEX forma parte del **Consejo Interterritorial de Internacionalización**, en el que también participan las Comunidades Autónomas, el Consejo Superior de Cámaras y la Confederación Española de Organizaciones Empresariales (CEOE).

El ICEX participa como accionista en las empresas públicas:

- P4R: especializada en proyectos en países menos desarrollados.
- COFIDES: ofrece financiación para proyectos privados viables en el exterior en los que exista algún tipo de interés español.

⁷⁵ <http://www.icex.es/icex/es/index.html>

7.3.3.2. Red de Cámaras de Comercio de España en el Exterior (CAMACOES)

Es uno de los instrumentos más eficientes y que más puede ayudar en la búsqueda de los objetivos. Cuenta con una **amplia red de sucursales en numerosos países**⁷⁶ los cuales se hacen eco de las necesidades e inquietudes facilitando los protocolos y procesos necesarios para la expansión a realizar. Como ejemplo, se detalla el siguiente:

- **Programa PIPE** (Plan de Iniciativa a la Promoción Exterior)
 - Es un programa conjunto de las **Cámaras de Comercio e ICEX**, en colaboración con las Comunidades Autónomas y la Unión Europea, que tiene como objetivo impulsar la internacionalización de las Pequeñas y Medianas empresas españolas y conseguir que se conviertan en nuevos exportadores. Ofrece apoyo y asesoramiento individual y especializado. Ofrece los servicios de diferentes profesionales para ayudar en todo el proceso de internacionalización:
 - **Tutores PIPE**: profesionales que normalmente pertenecen al ICEX o a las cámaras de comercio que asesoran a la empresa durante la primera fase del proceso.
 - **Promotores PIPE**: consultores externos que ayudan a la empresa en la confección de su Plan de internacionalización.
 - **Colaboradores PIPE**: jóvenes con formación sobre el plan que la empresa puede contratar para poner en marcha el Plan de internacionalización, así como realizar las actividades de promoción del mismo.



7.3.3.3. Instituto Cervantes

El Instituto Cervantes es la institución pública creada por España en 1991 para la promoción y la enseñanza de la **lengua española y de las lenguas cooficiales** y para la **difusión de la cultura española e hispanoamericana**. Está presente en 90 ciudades de 43 países en los cinco continentes⁷⁷.

- **Objetivos y funciones**
 - Organizar **cursos generales y especiales** de lengua española, así como de las lenguas cooficiales en España.
 - Apoyar la **labor de los hispanistas**.
 - Participar en **programas de difusión** de la lengua española.

El Instituto Cervantes cuenta con dos sedes en España, la sede central de Madrid y la sede de Alcalá de Henares, y, dentro de su plataforma multimedia, cuenta con un portal de referencia en Internet sobre la lengua española y la cultura en español: el Centro Virtual Cervantes.

⁷⁶ <http://www.comercio.gob.es/es-es/comercio-exterior/informacion-sobre-paises/red-exterior/red-camaras-oficiales/paginas/camaras.aspx>

⁷⁷ www.cervantes.es

- Realizar actividades de **difusión cultural** en colaboración con otros organismos españoles e hispanoamericanos y con entidades de los países anfitriones.
- Poner a disposición del público, **bibliotecas provistas de los medios tecnológicos más avanzados.**
- **Difusión de la lengua Española.** El Instituto Cervantes ha estado presente desde su fundación en diferentes ferias de la lengua, de la educación y del libro por considerar que son una plataforma idónea para dar a conocer a un público amplio, y a la vez especializado, sus objetivos: la promoción de la enseñanza del español en todo el mundo y la difusión de la cultura en español.



7.3.3.4. Embajadas (oficinas comerciales).

Mediante la llamada «**Ventana Global**», una empresa con intenciones de internacionalizarse, contando con más de 100 oficinas en el exterior, puede encontrar asesores que le informarán y guiarán sobre:

- Potencial de la empresa.
- Trámites de exportación.
- Elaboración de la estrategia internacional.
- precios de alquileres/ salarios.
- Información sobre posibles socios locales.
- Servicios de consultoría adaptados a las necesidades de la empresa.
- Esfuerzo financiero. Se deben conocer todas las puertas dónde se puede llamar.



En la Tabla 2 del Anexo se detallan las empresas españolas destacadas de la industria de la lengua, los casos de uso de las Tecnologías del Lenguaje en las TIC y otros productos de PLN y TA en las TIC.

Casos de éxito de empresas españolas

La prioridad en cuanto a la cobertura de mercado suele ser el nacional, con una especialización en la cobertura de lenguas co-oficiales. En el caso de la empresa encuestada que ya ha empezado una estrategia de internacionalización (Imaxin), ha considerado el mercado iberoamericano y en particular Brasil como su objetivo comercial, con herramientas para gallego, portugués y brasileño como una secuencia natural por la proximidad de las lenguas.

Hay otras experiencias de internacionalización conocidas: Bitext, Verbio o Sigma son empresas con delegaciones en EEUU, todas ellas ofrecen herramientas multilingües con un elenco de idiomas que cubre en el caso de Bitext, por ejemplo: Inglés, Español, Francés, Portugués, Italiano, Alemán, Neerlandés además de Catalán, Vasco y Gallego.

Conclusiones

Seguidamente se enuncian las **principales conclusiones obtenidas** de las labores de investigación realizadas en esta sección de internacionalización:

- El número de empresas independientes que utilizan estas funcionalidades está **aumentando progresivamente**.
- A pesar de la crisis, **se está invirtiendo cada vez más en empresas “startups” relacionadas con las máquinas inteligentes**. Los gigantes informáticos (IBM, Microsoft, Facebook, etc.) ha centrado su atención en este tipo de empresas para acelerar sus procesos de desarrollo en estas funcionalidades o mejorar las herramientas que comercializan en analítica avanzada.
- Las máquinas inteligentes van a tener **un impacto significativo en todos los sectores** de la industria en los próximos cinco años:
 - En el año 2018 las máquinas inteligentes (máquinas que toman decisiones basadas en el auto-aprendizaje) **distribuirán el 10% del trabajo humano**.
 - En el año 2018 el **30% de las interacciones que se realicen con aplicaciones informáticas se realizan a través de máquinas inteligentes**.
 - En el año 2018 en algún país desarrollado se habrá creado **jurisprudencia/marco jurídico para las máquinas inteligentes**: posibilidad de fallo y responsabilidad penal.

En los datos recopilados sobre empresas (informe Indra) se aprecia que la mayoría de empresas españolas tienen productos potencialmente de cobertura mundial pero que pocas han empezado estrategias de internacionalización. Ver Anexo "Datos de empresas españolas (información recopilada por Indra).



- En el año 2020 el 40% de altos ejecutivos distribuirá su trabajo para **versiones virtuales de sí mismos**, y aumentar así su productividad personal.
- Se considera conveniente impulsar las tecnologías relacionadas con lo siguiente:
 - Búsqueda de documentos.
 - Recuperación de información.
 - Presentación eficiente de información.
 - Traducción automática.
 - Asistentes personales.
 - Tecnologías del habla.
 - Gestores de publicaciones.
 - Gestores de contenido.
 - Organización de procesos de publicación.
 - Entornos de publicación multilingüe en tiempo real.
- Se ha comprobado que, pese a su madurez, el **NLP es una disciplina viva y en pleno desarrollo**, con un número importante de retos que superar.
- Es muy importante el **diseño de una buena interface usable y accesible**.

De todo lo dicho, se deduce la necesidad de continuar trabajando con el fin de poder divulgar el tratamiento del español como lengua en las diferentes herramientas conformadas a tal efecto.

7.3.4. Análisis DAFO

A continuación, se expone un análisis DAFO sobre la internacionalización.

Oportunidades

- Europa necesita **infraestructuras de lenguaje común**.
- **500 millones de hispano hablantes**: gran mercado potencial.
- La **investigación en PLN** abarca casi todos los ámbitos.
- Existen **redes temáticas** sobre PLN en español.
- Existencia de **programas de doctorado** con temática en PLN.
- Existen **países europeos** con estudios de grado en PLN.

Fortalezas

- **Industria** madura con demanda mundial.
- Centros de **investigación** como vehículo de transmisión mundial
- **Centros de investigación** en consorcios industriales.



- **Instituciones Españolas** estandarizadoras del lenguaje.
- Existencia en España de **financiación**.
- Existencia en España de productos comerciales.

Debilidades

- **Fragmentación** elevada de las empresas
- **Barreras** de mercado en países objeto de expansión.
- **Poca financiación** con fondos públicos
- **Solapamiento** en la oferta de productos.
- No existen **alianzas estratégicas** en el sector.

Amenazas

- IBM, Google, Microsoft, Apple, Skype, Intel ofrecen **soluciones en muy poco tiempo**.
- Mentalidad de **no compartir** los recursos: desconfianza.
- Falta de **continuidad en la financiación** de la investigación.
- El **inglés** como lengua de referencia en Latinoamérica.
- Las **PYMES** europeas no crecen.



7.4. Análisis de la participación de industria TIC

El objetivo de esta sección es **presentar y caracterizar los principales casos de uso de las Tecnologías del lenguaje en el ámbito del sector español de las Tecnologías de la Información y las Comunicaciones**.

Para ello, la sección toma como punto de partida la **descripción del ecosistema TIC español en relación a la explotación de las tecnologías del lenguaje**, incluyendo tendencias macro del sector, identificando las empresas que han registrado patentes en el ámbito de las TL y las empresas que han participado activamente en proyectos de innovación en cooperación relacionados.

Finalmente, en base a estas fuentes de información, se presenta **las conclusiones en forma de DAFO** para la industria española y se indican algunas recomendaciones al respecto.

7.4.1. El lenguaje natural en la industria TIC española

De manera relevante, **la presencia de la lengua española en las TIC a escala global**, no corresponde, en la medida en que cabría esperar, a la presencia de las tecnologías de la lengua en las TIC españolas.

Es en Latinoamérica donde se encuentra la mayoría de los hablantes nativos de lengua castellana, y que, si bien la presencia de las empresas españolas en latinoamérica es notoria (propiciada por la proximidad cultural, y sobre todo idiomática), el desarrollo de la economía latinoamericana da lugar a empresas de nueva creación de base tecnológica con capacidad de competir globalmente en este terreno.

El motivo fundamental es una **tendencia creciente al desarrollo y utilización de herramientas** que operan de manera independiente del lenguaje utilizado, combinado con la globalización de la economía y en particular los desarrollos TIC disponibles en la Cloud.

No obstante, cabe identificar **cuatro grupos de actores diferenciados presentes en el tejido español**:

- Como **habilitadores transversales**, los proveedores de servicios TIC tienen un perfil poco especializado en las TL:
 - Infraestructura y servicios de comunicaciones, como los operadores de red.
 - Infraestructura y servicios IT, como proveedores de servicios de cloud.
- En el otro extremo del espectro se encontraría el ecosistema de industria, academia e investigación predominantemente centrado en las TL, fundamentalmente:
 - Tecnologías del lenguaje
 - Tecnologías del conocimiento
 - Disciplinas relacionadas con el estudio de las lenguas (Lingüística Computacional dentro de Lingüística Aplicada)
- En una **zona intermedia**, se encontrarían los actores especializados:
 - En el manejo de grandes volúmenes de datos (Big Data)
 - En integración y algoritmia centrada en la sintaxis y semántica relacionados con las aplicaciones prácticas de las TL
- Por último, existiría un **cuarto grupo de actores**, compuesto por los **proveedores de contenidos, plataformas y bases de datos de información textual**, y muy en concreto información multilingüe. Esta información es utilizada con herramientas estadísticas, y es fundamental para alimentar mecanismos de aprendizaje automático y la base de los sistemas de Big Data. En este grupo se incluirían las agencias públicas que publican contenidos en varias lenguas y determinados grupos privados como multinacionales.

Las empresas competidoras más importantes en el mercado son la española Indisys (que ha sido recientemente adquirida por la multinacional Intel, lo cual es un reflejo de la importancia que las tecnologías de PLN tienen en el contexto de asistentes virtuales); y la multinacional sueca Artificial Solutions, (que incorpora reconocimiento de voz también en sus productos).

A través de la interacción directa y en algunos casos por referencias a partir del contacto con el tejido empresarial español, **CDTI ha identificado algunas de las empresas que trabajan actualmente en esta actividad** y que suelen estar estrechamente relacionadas con los Centros de Investigación de su Comunidad Autónoma:

- **País Vasco:** Fundación Eleka, Elhuyar, Anboto Europe, Todoentradas, Centro Tecnológico Vicomtech, Mapalia Networks, IK4-IDEKO, Mondragón Lingua, Sherpa.



- **Cataluña:** Lucy Software, Verbio Technologies, INCYTA, Inbenta profesional Services, AutomaticTrans, Viclone virtual solutions, websays, Tekstum, sisu-labs, tauyou <language technology>, Artificial Solutions. Ateknea Solutions, Inbenta, ExpertSystem Ibérica.
- **Andalucía:** MTC Soft, Victor Translator, Indysys,
- **Madrid:** Daedalus, Celer soluciones, Seprotect multilingual Solutions, Habla Computing, Molino de Ideas, Bitext innovations, Paradigma Tecnológico, Mamvo Performance, Sigma Technologies, Linguaserve, Sentisis, Playence, Fonetic Solutions, Semantia Lab, Hermés Traducciones.
- **Galicia:** Cilenis, Imaxin.
- **Extremadura:** Oteara.
- **Valencia:**, Innovative Social Technologies, Pangeanic, Prompsit.

Por último cabe citar a las **empresas relacionadas con el ámbito editorial** como fuertes motivadoras de estas actividades, con empresas como, por ejemplo:

- Editorial Tirant Lo Blanc.
- Calamo y Cran.
- Comercial de ediciones SM.
- Aralia Editores.

Según la asociación norteamericana para la globalización y localización (GALA) el mercado mundial de los servicios de lenguajes ha crecido a un ritmo anual del 8% en los últimos 3 años, estimándose en 2014 en casi 40.000 millones de dólares según la consultora especializada en idiomas Common Sense Advisory.

7.4.2. Principales casos de uso de las TL en las TIC

A pesar de tratarse de una disciplina con un largo historial de aplicación, los ciclos de innovación dentro de la industria TIC son cortos y dan lugar a nuevos casos de uso de manera continua.

Un ejemplo significativo se puede encontrar en el auge de los SVPA (**Asistentes Personales Virtuales Inteligentes**, por sus siglas en inglés). Si en 2014 se produjo un súbito interés en las aplicaciones móviles SVPA, en 2015 la mayoría han sido adquiridas por grandes corporaciones TIC, hasta el punto en que alguna de ellas apenas ha estado disponible algunas semanas. Emu fue adquirida por Google, Yahoo compró Donna y Cue fue comprada por Apple, por mencionar sólo alguno de los casos más sonados.

Emu, una vez descargada al teléfono móvil, supervisaba las conversaciones escritas que mantenía el usuario, ofreciéndole sugerencias a medida que esta se desarrollaba. Si, por ejemplo, el usuario sugería ir al cine, Emu utilizaba las capacidades de localización para ubicar ambos usuarios, sugería una sala de cine cercana, y mostraba la cartelera en la franja horaria deseada, filtrando las películas compatibles con las agendas de ambos, y ofreciendo detalles como la visualización del tráiler. Una vez escogida la sesión, ayudaba a comprar las entradas y actualizaba de manera automática la agenda de los usuarios.



Cabe suponer que la **tecnología subyacente se potenciará** encontrándose a su vez casos de uso novedosos.

La mayoría de las aplicaciones conocidas de las TL en diferentes dominios y sectores sigue siendo vigente, e incluso en algunos casos viendo aumentado su uso: universalización de contenido en la red, software embebido y aplicaciones mediante la traducción automática multi-idioma. Existe una **necesidad creciente de universalización**, que se apalanca en la globalización del acceso a la información, y aumentando de manera exponencial en las webs y aplicaciones que son actualizadas frecuentemente, especialmente en el caso de fuentes no profesionales (contenidos generados por el usuario).

Marketing, publicidad y atención al cliente

Las empresas, en particular aquellas con orientación al cliente final (B2C) han encontrado una valiosa fuente de información en el contenido existente en **foros, redes sociales, blogs y webs de opinión**. En algunos casos, el acceso a la información debe hacerse filtrando ingentes cantidades de contenido desestructurado.

Por último las empresas han habilitado nuevos canales de **comunicación con los clientes** (webs, apps, etc.) en los que el tratamiento automático en los primeros niveles, permite descargar los centros de atención al cliente tradicionales. Es conocido el caso de AETNA cuyo asistente virtual redujo en un 30% la carga de su centro de atención tras sólo 5 meses de su implantación.

Los comentarios de productos y servicios y de reputación de marca permiten valorar su posicionamiento, pero combinados con análisis más genéricos permiten segmentar y dirigir campañas de comunicación y acciones de marketing.

Política y administración pública

Al igual que en el ámbito empresarial, en el ámbito político, la extracción automatizada de opiniones permite **identificar, resumir y comprender opiniones y sentimientos** en grandes cantidades de información textual desestructurada.

También por parte de las administraciones públicas se están **popularizando canales de comunicación** basados en el lenguaje natural, facilitando el acceso a servicios a la ciudadanía y optimizando recursos públicos.

Educación

El uso de las TL en educación discurre de manera paralela al **crecimiento del uso de las TICs**, por ejemplo, con la proliferación de los cursos on-line (MOOCs y SPOCs).

Si bien es cierto que la aplicación de las tecnologías del lenguaje al aprendizaje de un idioma no es nueva, las nuevas posibilidades que ofrecen, las hacen unas **herramientas mucho más completas y cercanas** a la experiencia de un instructor humano.

El análisis de opiniones proveniente de grandes cantidades de textos permite en ocasiones anticipar movimientos económicos influenciados por las opiniones.

Adicionalmente al aprendizaje de un idioma, las TL se aplican en ámbitos concretos para dar acceso universal a contenidos localizados, aumentando el público objetivo de la acción formativa.

Finanzas y bolsas de valores

El sector financiero y bancario hace uso, en cuanto a empresa, de **los recursos de las tecnologías del lenguaje mencionados al respecto del marketing**, centros de atención al cliente, detección de fraude, enrutamiento de mensajes, extracción de información.

Salud y ciencias biosanitarias

Es conocido que algunos servicios de salud norteamericanos están realizando **consultas al sistema de inteligencia artificial IBM Watson** para casos clínicos de los que no se dispone de diagnóstico.

Por otra parte, el análisis masivo de textos permite **extraer información** relativa a la propagación de enfermedades, parámetros que favorecen determinadas patologías o en general información social que utilizar estadísticamente para correlar con datos médicos o epidemiológicos.

Seguridad

Automatizar el análisis de la comunicación de un usuario, permite **alertar de posibles incidencias de seguridad** o servir de un primer nivel de filtro que habilite un posterior análisis en el marco de la interceptación legal de las telecomunicaciones.

Al igual que en el caso de la salud, el **análisis agregado de grandes volúmenes de textos** permite anticipar tendencias sociales que tengan implicaciones en la seguridad pública.

7.4.3. Proyección de las TL en las TIC

El principal modelo de negocio de las Tecnologías del Lenguaje (TL) se basa en la **automatización del Procesamiento del Lenguaje Natural (PLN)**. A continuación se indican algunas de las conclusiones del análisis que se ha realizado de las distintas áreas comprendidas en PLN:

- **La mayoría de las soluciones de TA y de PLN se ofrecen a través de APIs** que permiten a los usuarios consumir los servicios en base a las necesidades. Estos servicios son de fácil integración en el desarrollo de nuevas aplicaciones y ofrecen resultados inmediatos y una modalidad pago por uso con costes muy económicos. Por el contrario, las soluciones comerciales completas disponen de algunos inconvenientes ya que se tratan de soluciones poco ágiles (tiempos de desarrollo elevado), de alto coste y con alta dependencia del proveedor.



- **El desarrollo de una suite de PLN o de TA no está al alcance de cualquier empresa.** Aunque existen numerosas suites de PLN de código abierto, son necesarios equipos multidisciplinares de especialistas en la lengua y en TI. La mayoría de empresas que disponen de productos propios son empresas con larga trayectoria en el ámbito y todas disponen de colaboraciones con universidades y participación en proyectos de I+D+i
- Desde una perspectiva empresarial se ha identificado la **necesidad de uso de una infraestructura de Big Data** (ecosistema Hadoop) y de herramientas para tratar la gran cantidad de datos desestructurados y poder analizar grandes cantidades de información de manera distribuida y de un modo eficiente.

En los siguientes apartados se describen y analizan las distintas **áreas de PLN** y se indican ejemplos de los cuales se puede extraer información relevante con respecto a la proyección de las TL en las TIC.

Traducción automática (TA)

El principal modelo de negocio está basado en la **comercialización de sus herramientas para la prestación de servicios de TA** a las Administraciones Públicas Autonómicas.

Los servicios de TA se ofrecen actualmente **en modalidad SaaS** (Software as a Service) en el cual los usuarios (clientes) pagan en función de su uso. En el caso de las administraciones públicas se realizan implementaciones llave en mano en la infraestructura de la propia administración (Opentrad en la Xunta de Galicia).

Como ejemplo cabe destacar la empresa [Imaxin Software](#), que actualmente se encuentra trabajando con el departamento de computación de la universidad de Santiago de Compostela en la paralelización con herramientas Big Data de la tecnología empleada en la mayoría de sistemas de TA denominada Moses SMT con el fin de mejorar el rendimiento en la prestación de servicios en otros sectores como son la TA en medios de comunicación (periódicos digitales).

Extracción de información

En esta área se engloban las empresas que trabajan **con tecnologías capaces de extraer información de textos** para poder crear una vista estructurada de la información mediante simplificaciones de texto. Esas tecnologías son capaces de proporcionar funcionalidades como:

- Detección de entidades nombradas independientemente del lenguaje
- Clasificación de entidades en distintas categorías, por ejemplo: *Madrid à Ciudad, Juan à Persona, El Corte Inglés à Organización*
- Detectar el idioma de textos al vuelo

La mayoría de las soluciones analizadas se basan en la tecnología Moses Statistical Machine Translation (SMT). Se han identificado empresas especializadas en TA principalmente en las comunidades autónomas con idioma propio como Imaxin con su producto Opentrad en Galicia, o MondragonLingua en el País Vasco.



- Extracción de relaciones semánticas donde se identifican las relaciones entre entidades (por ejemplo saber que una ciudad está relacionada con una persona en un texto)

Actualmente son muchas las empresas que compiten en esta área proporcionando el servicio mediante APIs en la modalidad SaaS para su utilización en desarrollos de terceros. Cilenis y Daedalus son ejemplos de ello. **Las tecnologías empleadas son similares ya que utilizan recursos lingüísticos propios (tesauros) generados a través de otros recursos como la Wikipedia o la DBPedia.** Por su parte, Indra ha desarrollado en el marco de un proyecto CENIT un sistema para automatizar la generación de recursos lingüísticos utilizando la tecnología Big Data. Debido a la alta competitividad y bajo coste del consumo de los servicios proporcionados por las empresas mediante APIs, en pocas ocasiones compensa implementar una solución de este tipo llave en mano. Otros participantes en proyectos CENIT o CIEN son ATOS, ISOCO-Fractalia, HAVAS-MEDIA.

Daedalus es una empresa que destaca por amplia experiencia en la recuperación de información de redes sociales aplicada al mundo del espectáculo. También destaca la empresa bitext cuyo principal modelo de negocio está centrado en el análisis de redes sociales para la atención al cliente de empresas de telecomunicaciones.

Recuperación de información

El almacenamiento de **grandes cantidades de información para su posterior indexación** juega un papel muy importante en la recuperación de la información. El principal fin de esta subárea se trata de obtener la capacidad de indexar mucha información para proporcionar al usuario mecanismos de indexación y acceso a esa información. Los principales productos identificados en el mercado se basan en la identificación de términos relevantes en un conjunto de documentos y la recuperación de contenidos de redes sociales como Twitter.

Clasificación

Tiene como objetivo **clasificar términos y/o documentos** en categorías predefinidas o agruparlos por similitud. Numerosas empresas que disponen de suites de PLN ofrecen esta funcionalidad como es el caso de Cilenis o Daedalus.

Creación de recursos lingüísticos (tesauros)

Existen ámbitos que disponen de un **dominio específico de la lengua**. Para estos ámbitos es necesario la elaboración de recursos específicos o diccionarios que contienen una jerga de un campo de conocimiento denominados tesauros. La mayoría de empresas dedicadas al PLN tuvieron que crear recursos lingüísticos propios para el desarrollo de sus productos, pero existen empresas como [Inbenta](#) que llevan años dedicándose a la realización de tesauros de ámbitos especializados como puede ser el sector jurídico.

Resumen automático de textos

En la actualidad, la tecnología **permite realizar resúmenes automáticos de textos asignando puntuaciones a las oraciones dentro de un documento**. Las oraciones que contienen palabras que se utilizan con más frecuencia en el documento se les da una puntuación más alta. Se trata de una funcionalidad importante en el desarrollo de numerosas aplicaciones, pero sencilla y que

actualmente proporcionan casi todas las empresas que disponen de sus propios productos de PLN.

Análisis de opinión y sentimientos

Actualmente es habitual **analizar la opinión de los usuarios en diversas redes sociales** como Twitter y de esa manera poder relacionar el sentimiento (positivo o negativo) de un usuario con respecto a una marca, un objeto, un lugar etc. Las empresas destacadas en recuperación de información disponen de tecnología de análisis de opinión y sentimientos con el fin de poder extraer esa información añadida de grandes cantidades de información indexadas por los sistemas de captura de datos. **Numerosas empresas se encuentran trabajando en el desarrollo de infraestructuras Big Data** para mejorar en rendimiento y velocidad el análisis de grandes cantidades de información.

Sistemas de diálogos

Estos sistemas reciben como **entrada frases del lenguaje natural y generan como salida frases del lenguaje natural**. Se utilizan en diversas empresas para proporcionar información de manera automática en asistentes virtuales, por ejemplo, horarios de salida de aviones, partes meteorológicos, estado de cuentas bancarias, etc.

Los sistemas de diálogos flexibles no se encuentran integrados en las principales suites de PLN quizás por tratarse de soluciones que deben tratar un lenguaje de un ámbito muy específico y ser poco transversal para comercializar de manera general el servicio. La solución genérica más completa conocida actualmente en el mercado es Watson desarrollada por IBM la cual utiliza las tecnologías más modernas sobre las que todos los competidores están trabajando actualmente como Big Data. IBM adquirió la empresa líder en el mercado de comercialización de servicios de PLN llamada AlchemyAPI cuyas funcionalidades fueron integradas en Watson. IBM ofrece todas las funcionalidades de esta plataforma a través del entorno de cloud computing Bluemix.



7.4.4. Tendencias del sector TIC en relación con las TL

A continuación se describen **tanto macro tendencias del sector TIC** que tienen particular aplicación a las TL como **tendencias concretas ligadas al procesamiento del lenguaje natural**.

Enfoques basados en tecnologías abiertas y creación de ecosistemas

El sector TIC europeo, en general, y el español, en particular, trabaja activamente en alinearse de modo que puedan ofrecer **competencia a las soluciones monolíticas y cerradas de los proveedores predominantes**, mediante la estandarización, adopción de APIs y tecnologías abiertas, el open source y en

Cabe destacar que muchas de las soluciones de software comercial, siendo muy avanzadas, apenas ofrecen características que las hagan adelantarse al software abierto.

general sistemas que favorezcan participación colaborativa y la creación de ecosistemas.

Servicios masivos disponibles en *Cloud*

Una tendencia masiva es la disponibilización de aplicaciones y servicios en paradigmas de Cloud Computing. Esto permite pasar de manera cada vez más sencilla de soluciones integradas en sistemas propietarios **a servicios bajo demanda o “llave en mano”**. Así como un despliegue más sencillo y estandarizado de soluciones de terceros.

Servicios como el mencionado IBM Watson, ha sido convertido en un servicio basado en la cloud a principios de año.

Aumento exponencial de la cantidad de información disponible

El llamado “Big Data” referido a las cantidades ingentes de información que se generan y procesan de manera continuada, aplica especialmente a las Tecnologías del Lenguaje, siendo estas **una de las principales estructuras portadoras de esta información**. Esto ha permitido, por ejemplo, un acercamiento a la traducción automática distinto al propugnado tradicionalmente con enfoque puramente sintáctico-semántico.

El llamado **enfoque estadístico** procesa el texto original buscando patrones similares en bases de datos de textos ingentes ya traducidos para aplicar después el patrón de traducción similar. Permite prescindir por tanto de gran parte del conocimiento sintáctico.

La **combinación de ambos métodos** (el enfoque híbrido) está dando resultados muy prometedores.

Tecnología multilinguaje o agnóstica del lenguaje

Tanto por la vocación global de los desarrollos TIC como por el modelo agnóstico del lenguaje del enfoque estadístico, se tiende a **desarrollar y utilizar en menor medida sistemas con una especialización en una lengua, dialecto o ámbito de aplicación concreto**.

Mayor interés en aspectos semánticos y contextualización

La relativa pérdida de relevancia de aspectos sintácticos frente a los análisis estadísticos ha **desplazado el foco hacia desarrollos que favorezcan la contextualización** y por tanto enriquezcan la capacidad semántica de los sistemas.

Sistemas para la elaboración y publicación de contenido con metadatos estructurales y semánticos

La eficacia de estos métodos va vinculada a la disposición de cantidades masivas de texto, y por tanto el funcionamiento en lenguas minoritarias o lenguas de las que se dispone de pocas traducciones es escasa.



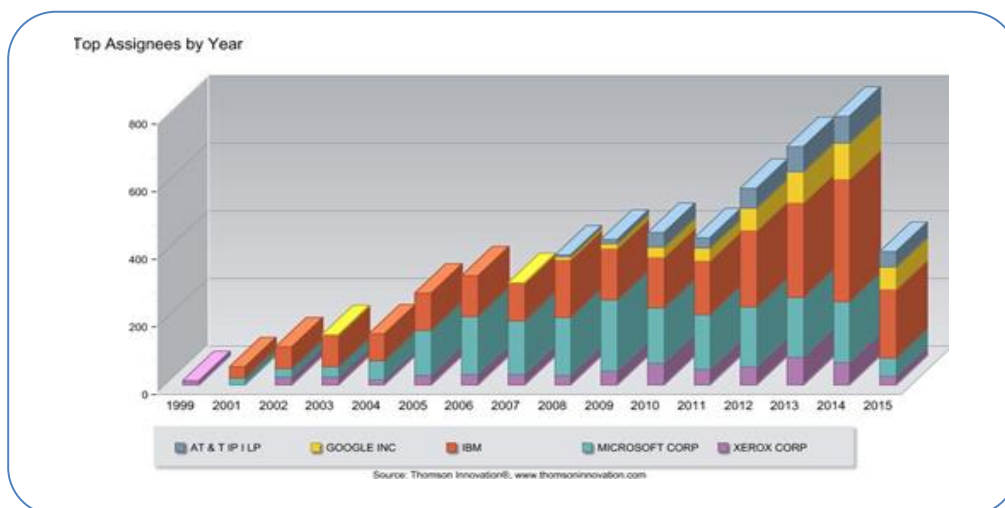
El creciente interés en maximizar el impacto de las webs, desarrollos, etc. está propiciando la **aparición de publicaciones cada vez más ricas en metadatos**, tanto estructurales como semánticos, con el objeto de facilitar el tratamiento automatizado de los textos producidos y su traducción.

7.4.5. Patentes relacionadas con el procesamiento del lenguaje natural. Especial mención de las patentes de empresas españolas.

El análisis en la evolución de las patentes a lo largo del tiempo y los principales cambios en sus actores (ver *el siguiente gráfico*) nos muestra que IBM, además de ser el principal titular, **ha incrementado de manera notable su implicación en la generación de tecnología basada en el lenguaje natural**. Por otro lado, se puede observar cómo Google está acelerando el número de sus patentes de manera continuada desde aproximadamente el año 2010, lo que muestra una tendencia a transferir la protección de algoritmos de búsqueda del know-how hacia las patentes.

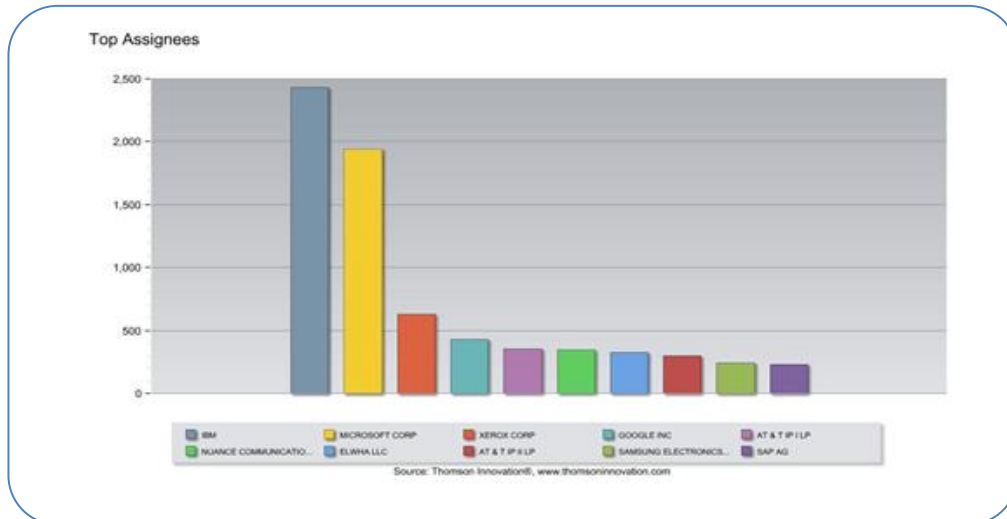


La tecnología patentada en relación con el lenguaje natural se concentra en pocas empresas, tal y como se puede observar en el gráfico siguiente. Entre ellas, cabe destacar **los grandes actores de este sector como IBM y Microsoft, la aparición de Google** en la generación de este tipo de tecnología, la presencia de Nuance con activos basados en el análisis de voz y su posterior análisis semántico, o la más bien ya histórica implicación de Xerox.



Principales titulares de patentes basadas en el lenguaje natural

La compañía Xerox mantiene una actividad constante en cuanto a nuevas patentes a un nivel medio o bajo. La empresa muestra una actividad comercial relevante en cuanto a la monetización o licenciamiento de su portafolio. Ese énfasis en la comercialización puede ser indicador de una menor actividad tecnológica o incipiente salida del sector del lenguaje natural.



Actividad en patentes por solicitante en el tiempo.

Con respecto al contenido de las patentes, cabe destacar que **la tecnología no es específica para un idioma concreto**. Ciertamente es que en muchas aplicaciones existe la necesidad de detectar rápidamente el idioma usado antes de pasar a un análisis más profundo de su contenido. En estos casos, el español es citado ampliamente en los documentos de patentes, pero casi siempre a título de ejemplo.



Por tanto, se llega a la conclusión de que **no existe una tendencia tecnológica específica para el castellano**, pero sí es un idioma elegido frecuentemente en tecnologías más genéricas y que tienen en cuenta varios idiomas. A continuación, mostramos las principales tendencias:

- Identificación de la **lengua utilizada en una comunicación** en tiempo real, como paso previo a un mejor tratamiento de la información del lenguaje natural.
- **Módulos de similitud** entre lenguas.
- **Módulos específicos por región o por idioma** que facilitan el procesamiento del lenguaje natural.
- **Uso del idioma como filtro** adicional en búsquedas de información y en la presentación de información (Por ejemplo en el Ranking de presentación de la búsqueda en línea).

Persiste la tendencia a proteger la tecnología del lenguaje natural como know-how. A modo de ejemplo, se pueden encontrar ofertas públicas de transferencia de know-how de varias universidades españolas.

En relación a la actividad de las empresas y universidades españolas, se puede decir que **éstas son más bien modestas**. Destacan las patentes relativas al asistente de voz Sherpa, cuya protección ha sido extendida a EEUU y Europa. Muchas de las patentes de otros titulares no están vigentes. En general, su tecnología o aplicación sería en los campos siguientes:

- **Sistemas de Respuesta de Voz Interactiva** o sistemas IVR (del inglés Interactive Voice Response) Uribe-Etxebarria (CEO del asistente virtual Sherpa), Progescar 2000 S A, Solaiemes S L, Inbenta y Retevision Movil, S.A.
- **Patrones y clusterización** semánticas del lenguaje natural: Universidad Madrid Carlos III, Universidad Catalunya Politécnica.
- **Aplicación a la Sincronización** de audio con subtítulos Universidad Madrid Carlos III.
- **Protección de contenido y conectividad segura** Forware Spain, S.L.

7.4.6. Participación española en proyectos subvencionados en el ámbito de las TL

Una de las herramientas que permite medir la presencia del empresariado español en el ámbito de las TL así como identificar a los actores del panorama nacional más activos, es **analizar el nivel de participación en las diferentes convocatorias de ayudas públicas a la I+D+I**, ya sean de ámbito nacional o internacional (fundamentalmente europeos).

A continuación se relacionan las entidades españolas que han sido más activas en las últimas convocatorias de proyectos de Investigación, Desarrollo e Innovación financiados por la Comisión Europea en el ámbito de las Tecnologías del Lenguaje:

- En convocatorias del séptimo programa marco de Investigación y Desarrollo (FP7)
- En las convocatorias del programa marco de Competitividad e Innovación (CIP)
- En las convocatorias de Horizonte 2020 centradas en Big Data y TL

Se han identificado **73 entidades españolas**, entre las que las siguientes participan en al menos dos proyectos diferentes⁷⁸:

- Universidad Politécnica de Madrid (7)
- Universidad del País Vasco, Ehu Upv (6)
- Atos Spain Sa (5)



Los informes de control de retorno de CDTI del octavo programa marco, ofrecen información desglosada por programa.

⁷⁸ http://cordis.europa.eu/projects/home_en.html

- Inmark Estudios Y Estrategias S.A. (5)
- Intelligent Software Components S.A. (5)
- Telefonica Investigación y Desarrollo S.A. (5)
- Universitat Pompeu Fabra (5)
- Fundació Barcelona Media, Universitat Pompeu Fabra (4)
- Universidad de Alcalá (4)
- Universitat Politècnica de Catalunya (4)
- Fundación Centro de Tecnologías de Interacción Visual y Comunicaciones Vicomtech (3)
- Fundación Tecnalía Research & Innovation (3)
- Paradigma Tecnológico SI (2)
- Universitat Politècnica de Valencia (2)
- Yahoo Iberia SI (2)

En el caso del **área 04.03 - (7ICT) TECNOLOGÍAS LINGÜÍSTICAS**, se presentaron **77 entidades españolas de un total de 770 entidades europeas**⁷⁹, resultando en 23 socios siendo financiados en un total de 39 proyectos, 9 de los cuales eran coordinados por entidades españolas. Las entidades más destacadas fueron:

- Universidad Politécnica de Cataluña
- Inmark Estudios y Estrategias, S.A.
- Universidad Politécnica de Madrid
- Universidad Politécnica de Valencia
- Universidad Pompeu Fabra
- Intelligent Software Components SA (ISOCO)
- Fundacio Barcelona Media (Universidad Pompeu Fabra)
- Asociación Centro De Tecnologías De Interacción
- Visual y Comunicaciones
- Atos Spain SA
- Universidad Del País Vasco / Euskal Herriko
- Unibertsitatea



En el caso del **área 04.02 - (8ICT) TECNOLOGÍAS DEL LENGUAJE**, hubo **propuestas de 20 entidades españolas diferentes de un total de 159**

⁷⁹ http://cordis.europa.eu/projects/home_en.html

entidades europeas, si bien únicamente ZABALA INNOVATION CONSULTING SA fue subvencionada, en un único proyecto en el que actuaba de coordinador⁸⁰.

En cuanto a las ayudas de la **SEIDI**, en los Retos Colaboración de 2014, se identificaron diez empresas participantes de siete proyectos relacionados:

- Ica Informatica y Comunicaciones Avanzadas SL
- Digital Preservation Solutions SL
- Tecnología e Ingeniería Sistemas y Servicios Avanzados De Telecomunicaciones, S.A.
- Nerlaska SL
- Invita Networks S.L.
- Zed Worldwide SA
- ScytI Secure Electronic Voting, S.A.
- Lks S.Coop.Ltda.
- Cognicor Technologies S.L.
- Lavinia Interactiva



Por último, destacar el proyecto del programa Innterconecta CELTIC, liderado por Indra Software Labs, S.L. en el que participan asimismo Elogia Media SL, Factoria De Software e Multimedia SL y SAEC Data SA.

7.4.7. Casos de éxito de la industria española TIC en el ámbito de las TL

Sigma Technologies

Sigma Technologies es una empresa tecnológica que desarrolla su actividad en las **áreas del procesamiento de señal y del reconocimiento de patrones**. El procesamiento de señal engloba una serie de algoritmos matemáticos que permiten extraer información y propiedades de las señales. Las técnicas de reconocimiento de patrones procesan y “aprenden” dicha información, de manera que posteriormente pueden identificar, de forma automática, distintos eventos, objetos y/o personas.

Los principales clientes de Sigma Technologies son grandes empresas del Silicon Valley, por lo que en el mes de julio de este año se está poniendo en marcha la oficina de Sigma Technologies en San Francisco.

En este sentido, en Sigma Technologies se trabaja en dos líneas de productos:

- **SigmaIdentity y SigmaInsight**. SigmaIdentity emplea las técnicas anteriormente mencionadas para verificar la identidad de una persona o para identificar a dicha persona por medio de distintos rasgos fisiológicos y de comportamiento. SigmaInsight extrae información de fuentes de

⁸⁰ http://cordis.europa.eu/projects/home_en.html

información desestructurada en formato de texto, de audio, de imágenes o de video.

La plataforma tecnológica que soporta SigmaIdentity y SigmaInsight está en **constante evolución**, de manera que, cada vez, va disponiendo de más funcionalidades y capacidades. Actualmente, la plataforma permite verificar la identidad de las personas o identificar personas a partir de su firma manuscrita y en breve podrá hacer lo mismo a partir de su voz. Asimismo, **se trabaja en cerca de 50 idiomas distintos**, con lo que la plataforma puede extraer información de textos en estos idiomas de forma automática. También transcribe e indexa automáticamente contenidos de audio y audiovisuales en español e inglés, lo que permite recuperar información de contenidos que actualmente sólo se pueden tratar de forma manual o semiautomática. Finalmente, la plataforma dispone de **herramientas que permiten procesar y obtener información a partir de imágenes y videos**. En este momento el sistema permite detectar y hacer seguimiento de personas y detectar objetos de forma automática.

Los principales clientes de Sigma Technologies están, actualmente, en Estados Unidos en las áreas de Internet y Telecomunicación y se está empezando a entrar en los sectores de healthcare, media & entertainment, financial services, defense y homeland security.

En particular, en el área de las tecnologías del habla y del lenguaje, se han desarrollado capacidades en varias líneas:

- **Capacidad para poder desarrollar productos de procesado del lenguaje en más de 50 idiomas/dialectos distintos**, tanto en la vertiente del procesado del lenguaje en formato de texto como en reconocimiento de voz. Para ello, se dispone de herramientas desarrolladas por la empresa y de una base de datos de cerca de 1.000 profesionales multidisciplinares, nativos en unos 50 idiomas, con los que se trabaja en los distintos proyectos que realiza la compañía.
- **Metodologías para el desarrollo de los recursos lingüísticos y biométricos** que tengan la calidad necesaria para el desarrollo de nuevas tecnologías y para el entrenamiento de sistemas automáticos de procesado del lenguaje natural, traducción automática, analítica de textos, reconocimiento de voz y sistemas de autenticación basados en parámetros biométricos y de comportamiento.
- **Industrialización de los procesos de desarrollo:** Desarrollo de metodologías y algoritmos de aprendizaje máquina que permiten, de forma automática, generar sistemas de procesado y clasificación de la información en un nuevo idioma y para una nueva aplicación de forma eficiente tanto en tiempo como en recursos humanos. Estas metodologías y algoritmos desarrollados permiten, por ejemplo, entrenar un clasificador de información en un nuevo idioma en tres o cuatro semanas, o desarrollar un sistema de reconocimiento de voz de habla natural en un nuevo idioma



en dos o tres meses, incluyendo, en estos plazos, el tiempo necesario para adquirir el volumen de datos necesario para entrenar estos sistemas.

- **Desarrollo de algoritmos adaptativos que permiten que un sistema ya definido** continúe aprendiendo el idioma para el que ha sido desarrollado, incluyendo las nuevas palabras y expresiones que vayan apareciendo.
- **Desarrollo de una plataforma que permite la combinación de distintas tecnologías** para la resolución de problemas complejos que no pueden ser abordados con un enfoque tradicional basado en una única fuente de conocimiento.
- **Know-how en la generación automática de resúmenes**, en analítica de textos y en traducción automática, aunque todavía no se dispone de productos en esta línea (están en fase de desarrollo).

Inbenta

Inbenta es una compañía española dedicada al **desarrollo de tecnologías relacionadas con el procesamiento del lenguaje natural y la búsqueda semántica** apoyada en un modelo de venta de sus productos basado en licencias. Con oficinas en diferentes países del mundo, enfoca sus esfuerzos en desarrollar técnicas de lingüística computacional, tecnologías Web y estrategias de marketing online.

A partir de su tecnología patentada, su principal objetivo es el de **mejorar la calidad de la atención al cliente al tiempo que se reducen los costes para las empresas**. Se trata, en definitiva, de mejorar la atención al usuario facilitándole respuestas rápidas y relevantes consiguiendo familiarizar al cliente con los contenidos de la web a la vez que se obtiene información crítica de las necesidades reales de los usuarios para aprender de ellas y mejorar de forma sostenida los contenidos.

A día de hoy estas “máquinas” son **capaces de detectar y clasificar de forma automática si las opiniones e interacciones en la experiencia de los clientes son positivas o negativas** (Inbenta Sentiment Detector), lo que resulta especialmente útil para aquellas corporaciones que necesitan mejorar la estrategia en sus servicios de atención al cliente y su reputación corporativa. Otra funcionalidad es **la retención cognitiva y de memoria que pueden llegar a tener los buscadores**. Por ejemplo, el Asistente Virtual de Iberia, la compañía aérea líder en su sector, es capaz de recordar el nombre del usuario, confirmando a las consultas un carácter personalizado, que aumenta la confianza en los resultados obtenidos y, por consiguiente, en la compañía.

En el año 2013 diferentes fondos de capital riesgo e inversores tecnológicos se interesaron por su tecnología, atendiendo a la rápida expansión. Fue en 2014 cuando la compañía consiguió levantar en una “Serie A” 2 millones de dólares con un fondo VC internacional, permitiéndole así continuar invirtiendo en I+D y posibilitando la creación de un software con sello español presente en todo el mundo. Actualmente Inbenta se encuentra en una **fase de constante**

La oferta de servicios y soluciones de Inbenta ha evolucionado en el tiempo hasta alcanzar las expectativas de grandes empresas, como Iberia, Groupon, Movistar, Casa del Libro, AXA Seguros, BBVA o Iberdrola entre otras, las cuales han reducido notablemente sus costes permitiendo al usuario llevar a cabo trámites y gestiones de manera autosuficiente, disminuyendo así las tediosas consultas y llamadas al Call Center correspondiente.



crecimiento, tanto a nivel nacional como internacional, cuadruplicando la facturación del primer trimestre del año anterior (2014)⁸¹.

Vicomtech

La Fundación Vicomtech-IK4, como centro de investigación aplicada y dentro de su departamento Tecnologías del Habla y del Procesamiento del Lenguaje Natural, ha participado y liderado diversos proyectos locales, nacionales e internacionales, **desarrollando y transfiriendo tecnología del lenguaje a la industria para la mejora de la productividad y la competitividad de las empresas de diferentes sectores**. Como casos de éxito de estos proyectos y de la tecnología desarrollada al servicio de las empresas, se pueden destacar los proyectos europeos SAVAS (Sharing AudioVisual contents for Automatic Subtitling), SUMAT (Online Service for Subtitling by Machine Translation) y OpeNER (Open Polarity Enhanced Name Entity Recognition).

Vicomtech-IK4 también lideró el **proyecto SUMAT** (www.sumat-project.eu), cuyo objetivo era el desarrollo y la integración de tecnologías de traducción automática en el flujo de trabajo de las empresas de subtitulación multilingüe. El proyecto se inició en relación con el **aumento continuo y drástico de los volúmenes de contenidos audiovisuales para los cuales se requiere una traducción de los subtítulos de alta calidad**. El desarrollo de tecnología de traducción automática basada en métodos estadísticos, y adaptada al dominio de la subtitulación, ha permitido aportar una respuesta a esta fuerte demanda, añadiendo una opción de mayor automatización a la traducción esencialmente manual empleada por la industria.

Basándose en las necesidades reales de la industria, se diseñó y desarrolló un **servicio en línea que permitiera la gestión de proyectos de traducción usando los sistemas de traducción automática de subtítulo desarrollados**. El proyecto incluyó dos amplias campañas de evaluación de la tecnología y de su uso por parte de usuarios profesionales del dominio. La primera evaluación se centró en la calidad de los sistemas creados en el marco del proyecto, resultando en un 67% de subtítulos determinados como requiriendo correcciones mínimas o ninguna corrección. La segunda evaluación se centró en medir los efectos del uso de la traducción automática sobre la productividad, resultando en una ganancia global alrededor del 40%. El proyecto ha tenido como impacto la adopción de la tecnología de traducción automática, inmediata o planificada, por parte de actores mayores del sector, ilustrando el impacto de las tecnologías de la lengua en un sector clave de la industria⁸².

El **proyecto OpeNER** (www.opener-project.org), también coordinado por Vicomtech-IK4, se centró en el **desarrollo de tecnologías abiertas para el tratamiento multilingüe de entidades nombradas y el análisis automático de sentimientos**. Estas tecnologías son de mayor importancia para la implantación

El proyecto reunió cuatro empresas especializadas en subtitulación, entre ellas dos multinacionales con fuerte presencia en el mercado, y cinco socios técnicos encargados del desarrollo de sistemas de traducción automática basados en datos profesionales, para nueve idiomas europeos combinados en catorce pares de idiomas-

⁸¹ <https://www.inbenta.com/en/newsletter/item/19>
⁸² (www.sumat-project.eu),

multilingüe y la expansión de mercado, y su potencial se comprobó para el dominio del turismo profesional en el marco del proyecto. A la vez, el desarrollo o la integración de las tecnologías básicas necesarias para el análisis de sentimientos y la minería de opiniones son costosas y pueden suponer en obstáculos para la creación de soluciones novedosas por parte de las PYMEs. Aportando una respuesta a este problema central para una óptima explotación de las tecnologías de la lengua, OpeNER ha compartido sus resultados y recursos bajo licencias de uso como Apache, favorables a usos comerciales. Con un enfoque basado en la aportación de tecnología de análisis multilingüe beneficiosa a la industria, una **cadena de recursos básicos esenciales se ha generado y compartido para el inglés, francés, alemán, holandés, italiano y castellano.**

Frente a mercados multilingües, particularmente en Europa con la emergencia de un mercado digital único cuyo multilingüismo es soportado por la mayoría de la población, las tecnologías de la lengua aportan respuestas que permiten superar barreras de entrada al mercado y generar ofertas multilingües a menor coste. Su impacto se ha demostrado con éxito en los proyectos presentados, representativos de la sinergia entre las tecnologías del lenguaje y las necesidades de las empresas en múltiples sectores.

7.4.8. Análisis DAFO

A continuación, se expone un análisis DAFO enfocado en el sector industrial español que está trabajando en TL.

Oportunidades

- El desarrollo y generalización de **tecnologías de Big Data**, o análisis de cantidades masivas de datos para extraer patrones y conclusiones, abre la puerta a nuevas aplicaciones para las tecnologías del lenguaje.
- El **crecimiento del español en el mundo**, y de las nuevas generaciones bien formadas en muchos de los países de habla española que son potenciales usuarios de este tipo de aplicaciones.
- Los **programas de I+D+I de la UE** que pueden ayudar a financiar nuevos proyectos de I+D en este campo.



Fortalezas

- El **auge creciente del español y la gran cantidad de textos disponibles**. La existencia de un sector TIC español potente que está demostrando su capacidad para competir globalmente. Un ejemplo sería su liderazgo en el desarrollo de plataformas sobre Internet.
- La **influencia del sector español hacia los países de habla hispana**.

Debilidades

- Se considera que el sector de empresas que trabajan en TL es **pequeño y no está suficiente organizado** para actuar de forma unificada. Además, algunas de las nuevas tecnologías de lenguaje natural son independientes del idioma; y por lo tanto, no es fácil utilizar la potencia del idioma castellano, como factor diferencial. Las empresas que trabajan con foco en las otras lenguas oficiales del Estado, no tienen una masa crítica de mercado potencial que les permite crecer.
- **El castellano, no es un idioma mayoritario en literatura científica**, patentes; por lo tanto existe una debilidad en cuanto al tratamiento de textos especializados.

Amenazas

- La principal amenaza para el sector español es la **tendencia creciente hacia herramientas y enfoques independientes del lenguaje**. Así grandes empresas como IBM, Google, SAS pueden desarrollar aplicaciones potentes adaptables a todos los idiomas.
- El **crecimiento del español en Estados Unidos**, si bien es una gran oportunidad en sí mismo, también implica el nacimiento de una competencia para las empresas españolas. Asimismo, el auge de start ups en países como Colombia, Perú, México, Chile, está también incrementando la competencia en este ámbito
- Desde el punto de vista de la aceptación de la tecnología, en el caso de las aplicaciones basadas en Big Data, el análisis de información puede ser **visto con recelo en cuanto a la privacidad de las comunicaciones**.

7.4.9. Recomendaciones

- **Crear conciencia sectorial.** El sector industrial español de empresas que trabajan en PLN no está suficientemente organizado. Para solucionarlo se propone promover alguna acción en torno a las asociaciones del sector (Ametec, entre otras). Asimismo, se propone promover la participación de las empresas en las asociaciones académicas de lenguaje natural.
- **Compra pública de tecnología.** Aprovechar las políticas de compra pública de tecnología en las adquisiciones de tecnología de los organismos públicos en aquellas aplicaciones que hacen un uso exhaustivo de las lenguas oficiales del Estado.
- **Incorporación de las tecnologías de LN en los programas de ayuda a la I+D promovidos por la AGE.** Incorporar en los programas de I+D del Ministerio de Industria, Energía y Turismo ayudas específicas a las pequeñas empresas de base tecnológica que trabajan en ese sector de las tecnologías sobre lenguaje natural.

- **Cooperación internacional.** Incorporar acciones de apoyo a este sector en los convenios de cooperación bilateral con los países de Latinoamérica. Se trataría de apoyar desde el sector tecnológico acciones en las que también tendrían cabida instituciones como el Instituto Cervantes, la Real Academia de la Lengua, etc.





8.

Las Administraciones
Públicas y el PLN y TA en
los servicios públicos

8. Las Administraciones Públicas y el PLN y TA en los servicios públicos

La administración ha realizado un **extraordinario esfuerzo en los últimos años para desplazar la tramitación desde los canales presenciales a canales telefónicos** y, principalmente, al canal de internet. En la medida de lo posible se han utilizado soluciones comunes, aunque la atomización de las unidades administrativas no siempre lo ha permitido. Además, ahora con un cambio necesario para proporcionar los mismos servicios en movilidad.

Existen gran cantidad de soluciones automatizadas de tramitación que, sin ser sofisticadas, se alejan de la tramitación tradicional.

Por otra parte, España es un país con **varias lenguas oficiales y una legislación que obliga a su traducción**, planteando como un derecho de los ciudadanos la posibilidad de poder realizar trámites con la administración en cualquiera de las lenguas oficiales de sus Comunidades Autónomas.

La crisis económica ha acelerado este desplazamiento y **condicionado las soluciones en varios sentidos**:

- La **tramitación electrónica** ya es, sin ninguna duda, la más utilizada por ciudadanos y empresas (éstas casi completamente de forma obligatoria).
- La creación de la **Dirección TIC** del Ministerio de Hacienda y Administraciones Públicas favorecerá la centralización tecnológica con el desarrollo de soluciones globales, comunes y horizontales.
- Los **desarrollos que se realizan son válidos** para todas las Administraciones Públicas, por lo que hay una masa crítica muy importante de organismos interesados.
- La necesidad de **traducción automatizada** es imprescindible al disminuir el presupuesto en todos los niveles, entre ellos, el dedicado a traducciones.

En este contexto, se deben **optimizar búsquedas, selecciones y mejorar la comunicación en general** con las nuevas opciones automáticas desarrolladas. La tecnología del lenguaje natural aplicada a las soluciones de las



administraciones públicas debe ser un campo prioritario de actuación por el beneficio general que conlleva.

La posibilidad de buscar en las sedes y portales, de seleccionar opciones de ayuda telefónica o en línea, utilizando el propio lenguaje del ciudadano, se plantea como un elemento crítico, en combinación con los buscadores clásicos del mercado para mejorar el servicio que se presta. El lenguaje formal es, fundamentalmente, utilizado entre los ordenadores para transferir información, pero es evidente la **necesidad de reconocer la sintaxis y la gramática**, la utilización de frases ambiguas o incompletas para mejorar la experiencia de usuario, optimizar la eficiencia de uso de los sistemas y realizar búsquedas complejas con fórmulas simples **en un ecosistema de administración electrónica cada vez con más opciones difíciles de seguir u organizar**.



8.1. Actuaciones en curso, necesidades detectadas y proyectos faro

En el contexto del grupo de trabajo se han definido **sectores verticales** (turismo, educación, sanidad) en los que se realizará un estudio intensivo. Se entiende que el sector de seguridad queda fuera de nuestro marco.

Así, el grupo de trabajo se ha centrado en el **trabajo clásico de gestión más administrativa** de las administraciones: tramitación de expedientes en el curso de un procedimiento, gestión de sedes electrónicas y otros portales electrónicos, relación con los ciudadanos y empresas por todos los canales, etc.

Como consecuencia de la definición de este marco de actuación, el grupo se ha planteado una serie de **necesidades específicas**:

- Mecanismos de **extensión a otras lenguas cooficiales**.
- Foco en las posibilidades que esta tecnología pueden aportar a las personas con **capacidades reducidas**.
- **Simplificación y mejora de la relación** de ciudadanos y empresas por cualquier canal y reducción de sus cargas administrativas.
- Determinación de los **proyectos FARO** en base a **maximizar el beneficio para la administración pública**.

En este marco, es importante resaltar que la **gestión administrativa**, aunque con un nivel de innovación tecnológica limitada, **es esencial para el funcionamiento** de las administraciones públicas. Afecta a todos los ciudadanos y empresas, con más de 500 millones de gestiones anuales por todos los canales (especialmente el canal telefónico y el acceso a internet) por los distintos dispositivos) con miles de millones de accesos a las páginas de internet. Por todo ello, cualquier mejora elemental en un **proceso general, por poco innovadora que pudiera parecer, tiene una trascendencia económica, estratégica y de mejora de la calidad de los servicios muy importante**.

Es evidente la necesidad de reconocer la sintaxis y la gramática, la utilización de frases ambiguas o incompletas para mejorar la experiencia de usuario.

Principales proyectos de la AGE en materia de procesamiento de lenguaje natural y traducción automática

A partir de la información contenida en la Plataforma de Contratación del Sector Público pueden obtenerse algunos **datos sobre los contratos públicos** en el ámbito de las tecnologías del lenguaje.

Los servicios contratados han sido principalmente:

- Sistemas automáticos de redirección de llamadas.
- Buscadores textuales avanzados.
- Desarrollo de ontologías y web semántica.
- Análisis de noticias, vigilancia sectorial.
- Traducción automática castellano-inglés, castellano-lenguas cooficiales.

Las comunidades autónomas con lenguas propias centran su interés en la **traducción automática y la generación de contenidos** en las lenguas cooficiales correspondientes.

Infraestructura de servicios de la red SARA

La Dirección de Tecnologías de la Información y las Comunicaciones perteneciente al Ministerio de Hacienda y Administraciones Públicas ofrece al resto de administraciones, a través de la red SARA, una serie de **servicios horizontales al resto de la Administración del Estado**.

En la tabla pueden verse los **principales servicios que provee esta infraestructura**. Desde el punto de vista de las tecnologías del lenguaje resultan especialmente relevantes los siguientes servicios:

Aplicación	Descripción
RCPDOC	Sistema de Gestión de Documentación Registral
BBDD-060	Sistema de información de legislación, becas y empleo
BUZON 060	Buzón 060
CISE	Centro de Interoperabilidad Semántica
DOCUMENTA	Almacén de Documentos de RRHH
FEP	Servicio Web de Fichas de Empleo Público
PORTAL060	Portal 060
PLATA	Plataforma de Traducción Automática
TRANSPARENCIA	Portal de implementación Ley de Transparencia



Las comunidades autónomas con lenguas propias centran su interés en la traducción automática y la generación de contenidos en las lenguas cooficiales correspondientes.

Destaca la existencia de la **Plataforma de traducción automática PLATA basada en motores de traducción open source**; Moses y Apertium. Moses es a su vez empleado por la plataforma europea MT@EU (base de la actual infraestructura de servicio digital (DSI) de traducción automática de la infraestructura digital Connecting Europe Facility, también llamada CEF).

PLATA permite la traducción automática de portales web a los idiomas cooficiales y al inglés a partir de textos en castellano. **El sistema permite la traducción al vuelo** de los contenidos de los portales o a través de servicios web. El sistema ofrece, cacheado de las traducciones, lo que repercute en el rendimiento de la plataforma.

El resto de servicios citados son **claros candidatos para la inclusión de tecnologías de procesamiento de lenguaje** avanzadas aplicadas a la mejora de buscadores, el encaminado inteligente de llamadas o la clasificación automática de documentos procesado al vuelo.

Conocimiento de la tecnología de PLN y TA por la administración

En el marco de la elaboración del presente informe ha sido enviado el **“Cuestionario sobre el uso de tecnologías del lenguaje 2015”** a las distintas Administraciones a través del Ministerio de la Presidencia y de la Dirección de Tecnologías de la Información y las Comunicaciones perteneciente al Ministerio de Hacienda y Administraciones Públicas. Dicha encuesta recoge distintos aspectos de las áreas de procesamiento de lenguaje natural y traducción automática. Por medio de esta encuesta se pretende:

- Recabar información sobre el **conocimiento y utilización** de este tipo de tecnologías.
- Conocer las principales **áreas de aplicación** (voz del consumidor, inteligencia competitiva, búsqueda, respuesta a preguntas, etc.)
- Conocer las **fuentes de información** sobre las que se aplican, la valoración económica de los proyectos y las lenguas en las que se realiza el procesamiento de lenguaje natural.
- Conocer el **tipo de procesamiento** así como las herramientas empleadas.

También es importante conocer cuál es el ámbito o dominio de los textos traducidos. Entre las opciones y necesidades detectadas se encuentran como **prioritarias**:

- **Traducción automatizada** de las sedes electrónicas y de los portales de páginas y documentos de las sedes, especialmente a lenguas cooficiales e inglés. Debería ser un servicio universal, masivo y prioritario. Esta necesidad tendría mucha sinergia con las iniciativas europeas de Connecting Europe Facility de traducción automatizada, un problema que, aunque general, aún no está suficientemente bien resuelto.



En el caso del área de traducción se ha consultado los pares de lenguas entre los que se realiza la traducción automática.



- **Reconocimiento automático de voz** en la gestión telefónica y procesamiento del lenguaje natural y utilización de motores de búsqueda semánticos. El desarrollo que se plantea en el presente y futuro de Apps hace que este campo sea también prioritario para las administraciones públicas.
- **Generación automática de voz a partir de texto** (call centers y elementos de usabilidad y accesibilidad).
- Reconocimiento automático de voz y procesamiento de lenguaje natural para páginas de ayuda, sistemas de **cita previa** y sondeos de calidad de servicios. También para la búsqueda en bases de datos de preguntas respuestas (FAQ, INFORMA, etc.)
- **Búsqueda inteligente en páginas webs, cada vez más complejas.** Es una necesidad evidente en un contexto de administración descentralizada en varios niveles como España, donde el ciudadano, en muchos casos, no conoce la administración competente.
- **Clasificación automatizada de documentos o precatalogación** de expedientes (ayudas, recursos, quejas y sugerencias, etc.).
- Servicios de **información administrativa** y de asistencia en problemas técnicos.
- Desarrollo de **asistentes virtuales**.
- Búsqueda y seguimiento de **patrones en páginas de internet** y redes sociales (seguridad, fraude, comercio electrónico, actividades no declaradas, etc.).

Otras áreas de posible aplicación son, por ejemplo, búsqueda de fraudes asociados a comercio electrónico o actividades no declaradas).

Algunos aspectos a considerar en el DAFO

- Ya existe un **sistema horizontal de traducción** en la nube (PLATA).
- Existe la posibilidad de **definir y completar data sets** en Dato.gob, específicos para estos desarrollos, seleccionado los catálogos y recursos existentes.
- Existe el **Centro de Interoperabilidad semántica, el Esquema Nacional de Interoperabilidad y del Centro de Transferencia de Tecnología**, elementos que pueden ayudar en la gobernanza, el trabajo cooperativo y el desarrollo de soluciones para todas las administraciones públicas.

La situación presupuestaria ha impedido trabajar en proyectos de futuro y valor añadido pero que no eran el “core” del negocio.

8.1.1. Análisis DAFO

Oportunidades

- **Necesidades horizontales** y globales para todas las AAPP.
- **Evolución mediante la DTIC** a elementos comunes para publicación de contenidos web en la AGE.
- Inclusión de colectivos con **limitaciones funcionales**.

- **Nueva generación de consultas web** mediante Apps.
- Generación de valor mediante la definición de **data set para datos.gob**.
- Existencia del **Centro de interoperabilidad semántica**.
- Existencia en el **CTT de forja** para permitir que las distintas administraciones definan soluciones comunes mejorables.

Fortalezas

- **Gran conocimiento del negocio** en sectores específicos (salud, seguridad, justicia, educación, turismo, hacienda, etc.)
- Existencia de **bases de datos termiológicas** (opciones, sinónimos, tesauros, topónimos, etc.) explotables globalmente.
- Clara determinación de las **necesidades prioritarias y masivas**.
- **Legislación básica** que obliga a que existan traducciones a lenguas cooficiales y soluciones técnicas de ayuda en el proceso.
- Existencia de algunos **desarrollos horizontales** que pueden servir de base (ej. PLATA)

Debilidades

- Todos los proyectos presentan **limitaciones para su implantación** completa.
- No existencia de una **norma específica de interoperabilidad**.
- No existencia de este tipo de soluciones aún en las **AAPP** para las aplicaciones móviles.
- Dificultad organizativa **de implicar a las CCAA y EELL** en tan escaso tiempo, en medio de un periodo de cambios.

Amenazas

- Las **traducciones deben tener validez jurídica** en las sedes electrónicas.
- La **situación presupuestaria** ha impedido trabajar en proyectos de futuro y valor añadido pero que no eran el “core” del negocio.
- **Pérdida de personal** especializado.
- Diferenciar entre el **procesamiento del lenguaje** y lo que debería ser lenguaje formal formateado XML, etc. (ej. búsquedas en el BOE)



8.1.2. Medidas propuestas

1. Plataforma común para el procesamiento de lenguaje natural y la traducción automática

Siguiendo el espíritu propuesto por la Comisión para la Reforma de las Administraciones Públicas (CORA), la Administración española se plantea la necesidad de la **eliminación de duplicidades administrativas** y de proporcionar servicios públicos comunes de calidad.

Se plantea, a continuación, una serie de áreas de aplicación, requisitos generales y particulares del servicio común que debe proporcionar capacidades de procesamiento de lenguaje natural avanzado y traducción automática. Dicha infraestructura será, en adelante, denominada como Infraestructura para las **Tecnologías Avanzadas de la Lengua (ITAL)**.

La infraestructura propuesta tiene un propósito múltiple:

- Aplicar de forma horizontal al conjunto de la Administración en su servicio al ciudadano
- Potenciar la investigación en el sector de las tecnologías del lenguaje
- Facilitar la innovación a pymes y emprendedores

Áreas de aplicación

Entre los proyectos más destacados por los expertos se encuentran:

- Traducción automática de lengua española, lenguas cooficiales e inglés.
- **Reconocimiento de voz y encaminamiento de llamadas en los servicios de Atención al Ciudadano**, adaptado al lenguaje administrativo y los dominios de información gestionados por la Administración.
- Búsqueda en **bases de datos documentales**. Catalogación y preclasificación de documentos.
- Asistencia en la **vigilancia sectorial**, visualización de grandes corpus documentales y detección de documentos similares.
- Asistentes virtuales en los **servicios electrónicos al ciudadano**.
- **Análisis** de texto, redes sociales y seguridad.

En cuanto a las áreas temáticas de aplicación son especialmente **estratégicas**:

- Sanidad y sector biotecnológico
- Educación y Cultura
- Turismo
- Justicia

Desde las primeras reuniones del Comité de Expertos que ha elaborado el presente informe se ha contado con la participación de la Oficina para la Ejecución de la Reforma de la Administración que ha planteado la necesidad de proporcionar servicios y tecnologías del lenguaje de forma común a toda la Administración.



- Seguridad y Defensa

Requisitos generales

La infraestructura común ITAL debe trabajar sobre las **lenguas oficiales de España** y facilitar la integración en el mercado único multilingüe europeo digital (MDSM) sirviendo como pasarela al resto de lenguas oficiales de la Comunidad Europea.

Dicha infraestructura debe ser especialmente sensible con el **uso por parte de personas discapacitadas de los servicios públicos**. Debe mejorar de forma inclusiva los servicios y la accesibilidad a los mismos.

La infraestructura debe dotar de servicios masivos basados en tecnologías del lenguaje a buena **parte de la Administración y proporcionar servicios horizontales sobre los que se apoyarán multitud de aplicaciones**. Por esta razón, la infraestructura debe facilitar *escalabilidad* horizontal y un crecimiento “elástico” (de forma transparente a las aplicaciones que los emplean) de todos los componentes que la conforman.

La integración de las aplicaciones puede variar desde las formas más sencillas por medio de servicios web a la integración en aplicaciones críticas en rendimiento o de reducidos recursos computacionales. La integración debe quedar garantizada por medio del uso **de estándares interoperables y componentes reutilizables**.

Asimismo, debe valorarse **la coexistencia de distintos modelos de licencias** para componentes de procesamiento y recursos lingüísticos tanto abiertos como privados. La infraestructura debe proporcionar mecanismos de contabilidad de uso que faciliten la incorporación a la misma de modelos de cobro por uso. La infraestructura debe facilitar múltiples modelos de negocio para la comercialización de recursos lingüísticos (procesadores y datos) así como corpus procesados.

Tanto en el modelo de uso como en el sistema de licencias empleado debe tenerse en cuenta desde un inicio el **múltiple propósito** de esta infraestructura: Administración, sector Investigador e Innovación. Dependiendo del papel de la infraestructura el modelo de procesamiento y licenciamiento puede ser distinto.

El presente informe, en el apartado 5, evalúa el **inmenso potencial de los recursos lingüísticos generados por la propia actividad pública**. Dicha iniciativa, llamada *Datos Abiertos de carácter lingüístico*, pretende generar todo tipo de recursos lingüísticos reutilizables por los componentes de esta infraestructura. Por esta razón, los estándares propuestos por dicha iniciativa deben ser soportados por esta infraestructura.

- El modelo descrito, debe facilitar la **inclusión de esta infraestructura**, con la posibilidad de tener *instancias separadas*, en tareas de:
 - **Investigación del propio** sector de las tecnologías del lenguaje

La diversidad de aplicaciones que conforman el ecosistema de la Administración Pública española hace necesario que los modelos de ejecución de dicha infraestructura tengan la posibilidad de dar servicio en la nube, de forma local en una administración dada (por razones de confidencialidad, rendimiento...) o de forma embebida en una aplicación existente.



- Aplicación en la **promoción de los proyectos de innovación** fomentados por la Administración.
- **Impulso a las PYMES** del sector y aquellas que quieran incorporar estas tecnologías como parte de sus productos o su proceso de producción.

En este sentido, la infraestructura debe actuar como una **incubadora de servicios** basados en tecnologías del lenguaje en su salto a la nube. No se encuentra entre los objetivos de ITAL proveer de servicios externos a las empresas, tan sólo impulsar su incorporación a la emergente explosión de aplicaciones innovadoras que todos los estudios pronostican para los próximos años.

La infraestructura debe permitir aportar contenidos y recursos propios para uso general o particular. Así mismo, debe facilitar compartir algunos de los corpus y subproductos generados de procesamiento (colecciones web descargadas, textos lematizado o con entidades reconocidas y clasificadas) para su uso en otras aplicaciones.

La infraestructura servirá como **ecosistema para la innovación**, facilitando la estandarización y evaluación de los desarrollos potenciados por la Administración tanto en el campo de la investigación como en el de la innovación.

La infraestructura **no pretende dar servicios completos** de aplicación que incluyan la visualización, la integración con interfaces de usuario o bases de datos de aplicaciones.

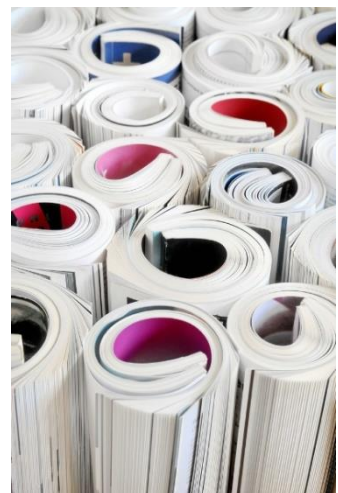
2. Requisitos de la infraestructura para el procesamiento de lenguaje natural

La infraestructura debe incorporar *recursos y procesadores lingüísticos interoperables y escalables*. Por ejemplo, la infraestructura facilitará la integración con marcos de trabajo conocidos: UIMA, GATE, Freeling, IXA pipeline, etc.

Permitirá la **anotación de documentos siguiendo estándares existentes** (NAF, GRAF, EAGLES-PAROLE, Universal Dependency TAGSET, CoNLL, Penn Treebank tag set, TIMEX, notación Stanford, etc.) y la incorporación de aquellos que se vayan imponiendo en el mercado. Así mismo, permitirá la inclusión de recursos lingüísticos siguiendo estándares que hagan interoperar los distintos componentes con diferentes conjuntos de datos y proveerá de conversores entre ellos.

La infraestructura debe **incorporar facilidades para la identificación de datos y herramientas**, la definición gráfica de los flujos de procesamiento lingüístico, la edición de las anotaciones, seguimiento de la ejecución de los trabajos, contabilidad de recursos consumidos, etc.

ITAL debe ser capaz de procesar tanto grandes corpus documentales como flujos de datos en tiempo real (twitts, noticias...). Así mismo debe permitir el procesamiento aislado de la información garantizando el control de acceso, el tratamiento adecuado de datos y la confidencialidad de la misma.



Requisitos de la infraestructura para la traducción automática

Tal y como se ha dicho anteriormente, ITAL debe trabajar sobre las lenguas oficiales de España y facilitar la integración en el mercado único multilingüe europeo digital (MDSM) sirviendo como pasarela al resto de idiomas oficiales de la Comunidad Europea. Además, dicha infraestructura debe contemplar las variantes del castellano empleadas en el resto del mundo, así como otras lenguas que puedan ser de interés para los distintos perfiles usuarios de la infraestructura.

Debe permitir el uso de **recursos lingüísticos**: corpus paralelos, reglas de traducción y diccionarios de traducción estándares tanto generales como especializados por dominios específicos y con diferentes modelos de licencias.

Los servicios de **traducción automática** deben proponer medios de Integración diversos, que variarán desde los proxy de servicios web para la integración inmediata en sitios web, adaptación a multitud de formatos estándar, etc.

La infraestructura debe **proporcionar capacidades de anonimización** de contenidos para la traducción de documentos protegidos por la Ley de Protección de Datos de carácter Personal (LOPD) o datos confidenciales.

La infraestructura debe seleccionar de forma semiautomática el **dominio de información al que pertenecen** los documentos del corpus documental a procesar por medio del análisis léxico, de tópicos u otras técnicas.

ITAL pondrá a disposición de usuarios especializados capacidades de postedición de textos traducidos y diccionarios, alineación semiautomática de textos, etc. Permitirá la anonimización empleando técnicas de reconocimiento de entidades.

8.2. Proyectos faro de las AA.PP. para impulsar la industria de las tecnologías del lenguaje

Los proyectos faro o bandera son proyectos emprendidos por las Administraciones Públicas de aplicación de las tecnologías del lenguaje natural en sectores estratégicos que pretenden servir de demostración de sus capacidades y beneficios, generar industria y generar recursos reutilizables en otros proyectos. También servirán de aprendizaje para futuros proyectos.

Los proyectos faro pretenden complementar las demás medidas del **Plan para el Impulso de la Industria de la tecnología del Lenguaje Natural**, que son de naturaleza horizontal, transversal, con actuaciones en sectores concretos que abarquen toda la cadena de valor y den lugar a productos y servicios acabados.

Existen numerosos ámbitos estratégicos donde las tecnologías del lenguaje podrían aportar grandes beneficios: Turismo, Educación, Sanidad, Justicia, Contenidos digitales, Medios de Comunicación, Videojuego, Comercio electrónico y exportación en sectores destacados de la industria española (textil –calzado, diseño, marroquinería, etc. agroalimentario –vino, jamón, aceite, etc. juguetería, ocio y espectáculos –fútbol, conciertos, etc.) etc.

Los **requisitos** para seleccionar los proyectos faros son los siguientes:



- **Sinergia** con sectores estratégicos de la economía o los servicios públicos españoles.
- Elevado **impacto** económico y social.
- **Desarrollo** de la industria de la tecnología del lenguaje nacional.
- **Concreción**.
- Desarrollo de la **cadena de valor** completa.
- Generación de **recursos reutilizables**.
- **Sinergias con las demás medidas del Plan** para el Impulso de la Industria de la Tecnología del Lenguaje Natural, y en particular, con la de generación de recursos lingüísticos (datos).
- **Compromiso** de los órganos competentes.
- **Carácter demostrador** de las capacidades y beneficios de las tecnologías del lenguaje.
- Especial atención a la adquisición de experiencia para futuros proyectos.

8.2.1. Análisis DAFO

Oportunidades

- **Existencia de sectores estratégicos** en lo que las tecnologías de lenguaje tendrían un gran efecto positivo.
- **Efecto demostrador** de los proyectos faro.
- **Aprendizaje** para proyectos futuros.
- **Desarrollo de la industria nacional** del lenguaje.
- Generación de **recursos reutilizables**.

Fortalezas

- **Potencialidad** del sector de las tecnologías del lenguaje español.
- **Sinergia con las demás medidas** del Plan para el Impulso de la Industria del Lenguaje Natural.

Debilidades

- **Desconocimiento** de los beneficios potenciales de las tecnologías del lenguaje.
- **Dificultad para segmentar** el proyecto en contratos no muy grandes adaptados a las características del sector español de la tecnología del lenguaje.

Amenazas

- **Compromiso insuficiente** por el órgano competente.
- Perder parte de la **cadena de valor**.
- Dificultades de **financiación**.
- **Absorción** por empresas extranjeras.

8.2.2. Medidas propuestas

Seleccionar y emprender un conjunto limitado **de proyectos faro de alto impacto**.

8.3. Datos lingüísticos abiertos: la AAPP proveedor de datos

Recursos lingüísticos generados por AA.PP. y la política de reutilización de la información del sector público (RISP)

La política de reutilización de la información del sector público (RISP) tiene como objetivo poner a disposición de la sociedad la información que el propio sector público genera en el ejercicio de sus funciones. Esto **persigue dos fines**: facilitar la transparencia y la rendición de cuentas de la acción gubernamental, y ofrecer de manera abierta un recurso que puede ser explotado económicamente.

A modo de **ejemplos de información generada por el sector público** de especial utilidad para el procesamiento del lenguaje natural se pueden citar:

Listas de nombres:

- Nombre de personas.
- Topónimos y direcciones.
- Directorios de organizaciones (empresas, administraciones, etc.).
- Marcas.
- Catálogos bibliográficos, filmográficos, del patrimonio cultural, titulaciones, etc.
- Diccionarios y glosarios especializados.

Corpórea de documentos disponibles en **más de una lengua**:

- Normas legales.
- Licitaciones.
- Subtítulos de películas.
- Notas de prensa y discursos.
- Convalidaciones de estudios.



La política RISP representa una extraordinaria oportunidad para el desarrollo de la industria del procesamiento del lenguaje natural por el gran valor potencial que tiene buena parte de la información que genera el sector público como recurso lingüístico.

Sirva para ilustrar la utilidad de la información generada por el sector público para el procesamiento del lenguaje natural el siguiente dato: en el portal de datos abiertos de la Unión Europea, **el conjunto de datos más visitado y descargado es la memoria de traducción de la Dirección General de Traducción de la Comisión Europea**. Se trata de la recopilación de materiales textos y traducciones llevadas a cabo por el personal de esta Dirección General, normalizados y estructurados para que puedan ser utilizados en la construcción de sistemas de traducción automática y como materiales para los sistemas de traducción asistida (<https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory>). Contiene textos, organizados en frases alineadas de 24 idiomas, formando un corpus con un total de 1.472.383.570 palabras.



En la **Unión Europea existe una política RISP común que pretende promoverse en toda la UE**, y hacerlo de manera coherente para facilitar el desarrollo de mercado único digital. Con este propósito se aprobó la Directiva 2003/98/CE del Parlamento Europeo y del Consejo de 17 de noviembre de 2003 relativa a la reutilización de la información del sector público. En España la RISP está regulada mediante la ley que transpone esta Directiva, la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público; y el reglamento que la desarrolla, publicado mediante el Real Decreto 1495/2011, de 24 de octubre, por el que se desarrolla la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, para el ámbito del sector público estatal.

La política RISP se apoya en el propio desarrollo de la sociedad de la información para facilitar el acceso a la información generada por el sector público, que se realiza en España mediante el portal <http://datos.gob.es/>

En el ámbito de la Administración General del Estado corresponde conjuntamente a la Secretaría de Estado de Administraciones Públicas, y la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información la promoción de la RISP. Con este fin, dentro del Plan de impulso de la economía digital y los contenidos digitales de la **Agenda Digital para España**, hay un programa de reutilización de la información del sector público. Este programa contiene medidas de promoción de la cultura RISP (iniciativa Aporta), para la consolidación y evolución del Catálogo de información pública, de asesoramiento y soporte, y de seguimiento y análisis.

Cabe señalar que, según datos publicados en la 3ª Conferencia Internacional de Datos Abiertos (2015, Ottawa, Canadá. <http://opendatacon.org/>), España es el **cuarto país del mundo en el clasificación de datos abiertos gubernamentales**.

Como ejemplos de informaciones del sector público con valor como recurso lingüístico que ya se pueden encontrar en el portal RISP español se pueden citar los siguientes publicados por la **Biblioteca Nacional de España**:

- Catálogo Bibliográfico de la Biblioteca Nacional
- Catálogo Colectivo de Publicaciones Periódicas (CCPP) de la Biblioteca Nacional
- Directorio de Bibliotecas Españolas



- Bibliografía española de publicaciones periódicas

8.3.1. Análisis DAFO

Oportunidades

- Desarrollo de la industria del **procesamiento del lenguaje natural**.

Fortalezas

- El sector público tiene **mucha información** susceptible de convertirse en recurso lingüístico.

Debilidades

- **Escasez de recursos** para facilitar la apertura de datos lingüísticos.
- **Escasa cultura RISP** tanto en la propia administración como en la sociedad.
- **Política de licencias y tarifas**.
- Acceso a los **datos abiertos lingüísticos**.

Amenazas

- **Normalización, calidad e interoperabilidad** de los formatos.
- **Expolio opaco** de recursos lingüísticos públicos.
- **Filtrado de datos personales**.

8.3.2. Medidas propuestas

- **Identificar y catalogar la información** del sector público susceptible de convertirse en recurso lingüístico.
- Adopción de **formatos normalizados**, de calidad e interoperables.
- Adopción de **técnicas de anonimización** adecuadas.
- Diseño de una política de **licencias y tarifas adecuada**.
- Facilitar el **acceso a los recursos lingüísticos** del sector público identificados mediante la **política RISP** en los formatos que faciliten su explotación y una experiencia de usuario avanzada.
- Fomentar la **cultura y el conocimiento** de los datos abiertos lingüísticos, tanto en las AA.PP. como en la sociedad.
- Participar en la **4ª Conferencia Internacional de Datos Abiertos (IODC)**, aprovechando que se celebrará en Madrid el 6 y 7 de octubre en Madrid, organizada por la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información, para dar visibilidad a la política de datos abiertos lingüísticos.



9.

Anexos

9. Anexos

Tabla 1: Cuadro resumen de empresas españolas de la industria de la lengua en España (experiencia recopilada por Indra)

Empresa	BiText	Cilenis	Daedalus	Expert System	Imaxin Software	Molino de ideas	Semantia Lab
Idiomas	Español	Español Inglés Portugués	Español	Español	Español Inglés Portugués Gallego	Español	Español
PLN	•	•	•	•	•	•	•
TA					•		
Identificación de idioma			•		•		
Extracción de <i>topics</i>	•		•		•		
Resúmenes de textos					•		
Clasificación de textos	•		•	•	•		•
Corrección de textos			•		•		
Lematización			•		•		
Análisis morfosintáctico			•		•		
Reputación corporativa			•		•		

Empresa	BiText	Cilenis	Daedalus	Expert System	Imaxin Software	Molino de ideas	Semantia Lab
Análisis de sentimientos	•			•			
Asistente virtual				•			
Personalización			•		•		
Normalización					•		
Opinión mining					•		
Otros		Herramientas lingüísticas Aprendizaje de idiomas		Desambiguación Etiquetado	Biomedical text mining Gamificación	Herramientas lingüísticas	Categorización Indexación
Mercado	Nacional	Nacional	Nacional	Internacional	Nacional Latinoamericano (principalmente Brasil) Europeo (principalmente Portugal)	Nacional	Nacional
Nº de empleados	Pyme < 10	Pyme	Pyme < 10		Pyme ~15		
Especialistas y doctores	Especialistas en NLP	<i>Spin-off</i> universitaria. Doctores en el ámbito de la especialidad	Especialistas en NLP que trabajan en la SEPLN	Especialistas en NLP	Especialistas en TA y NLP. Doctores que trabajan en proyectos en colaboración con OPIs	PLN (Doctores en el ámbito de la especialidad)	PLN (Doctores en el ámbito de la especialidad)
Tecnología propia o licenciada	Tecnología propia	Tecnología propia	Tecnología propia	Tecnología propia	Tecnología propia	Tecnología propia	Tecnología propia
Productos	Natural Opinions	CilenisAPI Linguakit Avalingua		Cógito	OpenTrad Galgo Exeria Coruxa Coati	WSO2 API Manager	Herramientas propias

Empresa	BiText	Cilenis	Daedalus	Expert System	Imaxin Software	Molino de ideas	Semantia Lab
Servicios. Modelo de negocio	Comercialización en el sector de teleco	Software libre para integración de PLN en software de terceros, herramientas lingüísticas a medida y aprendizaje de idiomas	Modelo de negocio SaaS Comercialización de APIs	Modelo de negocio SaaS Comercialización de APIs	Su principal producto se basa en la TA (OpenTrad)	Modelo de negocio SaaS	Modelo de negocio SaaS
Subsector			Social media	Banca Sanidad Energía Publicidad	Gamificación	Lingüística Formación	
Innovación. Líneas de investigación. Patentes		Gran componente investigador en todos los ámbitos del PLN en general			Big Data, Language Resources, Sentiment Analysis, Opinion Mining, Machine Translation, Online services using Open-source NLP tools, Computer Aided Language Learning		Empresa de base tecnológica con el respaldo académico de la UNED
Financiación pública anual		80% de sus ingresos se basa en financiación pública	Sí		60% de la facturación anual		
Proyectos nacionales o Europeos		Es una empresa joven que actualmente trabaja en proyectos nacionales	Participan en proyectos de I+D+i en convocatorias nacionales e internacionales		Participan en proyectos de I+D+i en convocatorias nacionales e internacionales		
Facturación							

Empresa	BiText	Cilenis	Daedalus	Expert System	Imaxin Software	Molino de ideas	Semantia Lab
Convenios con universidades		USC. Grupo de investigación ProINAT	Trabajan en proyectos con universidades y colaboran con diversos departamentos en el ámbito del PLN		USC. CITIUS Dpto. de lengua española (Dr. Paulo Gamallo) Dpto. de Ciencias de la Computación (Dr. Juan Carlos Pichel)		UNED
Formación en másteres. Doctorados		Sí					
Publicaciones					http://gramatica.usc.es/~gamallo/publis.htm		
Becas. Dirección de tesis		Sí	Beneficiaria de una ayuda del Programa Torres Quevedo para la contratación de un doctor				
Necesidades	Aplicación de tecnologías Big Data para procesos de PLN	Aplicación de tecnologías Big Data para procesos de PLN			Aplicación de tecnologías Big Data para procesos de TA y PLN		

Tabla 2: Casos de uso de las Tecnologías del Lenguaje en las TIC (experiencia recopilada por Indra)

Caso de uso	NLP, TA	Lenguas, Variantes, Dominios, Géneros	Contenidos digitales (Recuperación de información, Extracción de información, Simplificación de contenidos, Generación de resúmenes...)	Recursos y herramientas usadas	Necesidades y Demanda
Base de Datos de Traducciones de la Diputación Foral de Gipuzkoa. Banco Público de Memorias de Traducción del Gobierno Vasco.	TA	– Vasco	– Gestión de memorias de traducción. – No tiene restricciones en cuanto al número de lenguas, y se adapta a las necesidades particulares de cada cliente.	– eLENA , producto desarrollado por UZEI (Centro Vasco de Terminología y Lexicografía)	Gestión de memorias de traducción para diversos ámbitos. Generación corpus de traducciones multilingües y paralelos.
Verificador de léxico en vasco (Real Academia de la Lengua Vasca Euskaltzaindia, HABE, Diputación Foral de Gipuzkoa, Ayuntamiento de Donostia, Diario BERRIA, Editorial Elkar, Servicio de traducciones de la Universidad de Deusto)	NLP	– Vasco	– Verificador léxico en vasco	– IDITE , producto desarrollado por UZEI (Centro Vasco de Terminología y Lexicografía) – Basado en euLEX y UTH	
Traducción automática para las administraciones públicas del País Vasco	TA	– Español – Vasco	– Traducción automática de varios idiomas: Árabe, Vasco, Catalán, Chino, Gallego, Italiano, Japonés, Coreano, Rumano, Ruso, Sueco, Inglés, Francés, Alemán, Español, Portugués	– Elhuyar	Cubre la necesidad de traducción para diversos organismos del País Vasco
Traducción de documentos en la Administración pública de Galicia	TA	– Español – Gallego – Portugués	– Traduce entre diversas direcciones de traducción y también convierte textos entre diferentes variantes de algunas lenguas: Aragonés, Asturiano, Bretón, Catalán, Español, Francés, Galés, Gallego, Inglés, Italiano, Noruego, Occitano, Portugués	– OpenTrad de Imaxin – Apertium y Moses SMT	Traducción automática que también se comercializa como SaaS
Corrector ortográfico del Gallego integrado en Microsoft Word	NLP	– Gallego	– Corrector ortográfico para integrar en Microsoft Word	– Galgo de Imaxin	

Caso de uso	NLP, TA	Lenguas, Variantes, Dominios, Géneros	Contenidos digitales (Recuperación de información, Extracción de información, Simplificación de contenidos, Generación de resúmenes...)	Recursos y herramientas usadas	Necesidades y Demanda
Reconstrucción automática de operaciones de trading (sector financiero)	NLP	–	– Speech analytics	– ATR™ INTEGRA de Fonetic	Desbloquear el valor oculto que hay en las operaciones de trading y en las comunicaciones
Atención de clientes en sectores Telecom y Financiero	NLP	– 84 idiomas	– Speech analytics – Interactive Voice Response – Biometría de voz – Apto para diferentes acentos y dialectos	– Human-IVR™ de Fonetic – BIZ analytics™ de Fonetic – –	Reconocimiento de patrones de habla para estructurar y volcar al negocio la “voz del cliente”
Soluciones de traducción a medida para entornos corporativos	TA, NLP	–	– Traducción automática – Traductores híbridos libres para lenguas no emparentadas – Extractor y clasificador de información (Extractium) – Analizador de sentimiento (Opinum)	– Suite de productos Prompsit – Tecnologías Apertium y Moses SMT. –	Traductores híbridos basados en dos plataformas libres de traducción automática
Solución PLN ofrecida como API	NLP	– Inglés – Portugués – Español – Gallego	– Identificador de idioma – Análisis de sentimiento – Extracción de palabras clave – Reconocimiento de entidades	– CilenisAPI : conjunto de herramientas desarrolladas en el grupo de investigación ProINat@Ge de la Universidad de Santiago de Compostela.	Análisis de información
Solución PLN ofrecida como API	NLP	– Español	– Análisis de emociones – Lematización – Transcripciones fonéticas – Conjugador de verbos – Información léxica	– Apicultur de El molino de Ideas – Desarrollos lingüísticos propios	Análisis de información

9.3. Tabla 3: Distribución de las adscripciones en la actividad en PLN en América por países

Brasil (111)

- Interinstitutional Center for Research and Development in Computational Linguistics (NILC)
- University of São Paulo (USP)
- Federal University of São Carlos (UFSCar)
- State University of São Paulo (UNESP)
- Catholic University of Rio Grande do Sul (PUC-RS)

México (46)

- Fac. de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla (BUAP)
- Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (UNAM-IIMAS)
- Laboratorio del Lenguaje Natural, Centro de Investigación en Computación (CIC-IPN)
- Laboratorio de Tec. del Lenguaje, Inst. Nac. de Astrofísica Óptica y Electrónica (INAOE)
- Grupo de Ingeniería Lingüística (UNAM-GIL)
- ...

EE. UU. (18)

Argentina (12)

- Universidad Nacional de San Luis
- Universidad Nacional de Córdoba (GPLN-FaMAF)
- Universidad de Buenos Aires (DC-FCEyN)

Chile (10)

- Universidad Católica de Valparaíso
- Universidad de Concepción

Uruguay (8)

- Universidad de la República

Cuba (11)

- Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV) (6)

- Centro de Lingüística Aplicada de Santiago de Cuba (5)

Costa Rica (2)

- Universidad de Costa Rica

Venezuela (2)

Canadá (1)

Ecuador (1)

Tabla 4: Muestra de transferencia y colaboración universidad empresa

Nº	Grupo Universitario	Grupo	Proyecto	Empresa
1	Universidad de Murcia		Consultoría en anotación y recuperación en plataformas colaborativas	Edosoft factory, s.l.
			Consultoría en tecnologías de la web semántica y del procesamiento del lenguaje natural para la gestión del conocimiento en proyectos de desarrollo tecnológico.	
			Plataforma para la creación de asistentes virtuales basados en tecnología semántica	Vocali sistemas inteligentes s.l.
			Tecnologías de la Web Semántica para el turismo	
			Plataforma inteligente para gestión de ideas innovadoras open idea	Quality objects
			Sistema global para la gestión de la i+d+i en una organización (proyecto "iman")	Indra software labs s.l.u
2	Universidad Politécnica de Valencia (UPV)	Pattern Recongition and Human Language Technology	Desarrollo de un procesador de lenguaje natural	Contrato con Empresa
			Colaboración en el proyecto percepcion: desarrollo de sistemas de interacción avanzada hombre-maquina	Contrato con AETSI-Empresa
3	Universidad de Deusto	Morelab	Proyecto SELENE	LKS S. COOP.
			Traductor Automático para el Gobierno Vasco	Gobierno Vasco

Nº	Grupo Universitario	Grupo	Proyecto	Empresa
4	Universidad del País Vasco (UPV/EHU)	Grupo IXA	Lexikoaren Behatokia VII (2014)	Euskaltzaindia
			Lexikoaren Behatokia VI (2013)	Euskaltzaindia
			Lexikoaren Behatokia V (2012)	Euskaltzaindia

			Lexikoaren Behatokia IV (2011)	Euskaltzaindia
			Lexikoaren Behatokia III (2010)	Euskaltzaindia
			Lexikoaren Behatokia II (2009)	Euskaltzaindia
			ImFuTourNet-EHU (2009)	Elhuyar
			Lexikoaren Behatokia I (2008)	Euskaltzaindia
			Uztartzeko tresna informatikoaren diseinua eta integrazioa. (2007)	UPV/EHU y EJE, S.A.
			Multimeteo, generation in Basque of weather reports (2000)	Instituto Nacional de Meteorología , Lexiquet-France
5	Universidad Pompeu Fabra	TRL-IULATERM	Desarrollo de recursos para el análisis de expresiones lingüísticas en el seguimiento del "purchase funnel". (2013-2014)	Media Planning Group, Havas Media.
			Contrato de Investigación y Desarrollo (2013-2014)	Job and Talent, SL.

Tabla 5: Formación en NLP en España

Universidad	Grupo	Oferta	www
Universidad del País Vasco (UPV/EHU)	Facultad de Informática (Grupo IXA)	Máster en Tecnologías de la Lengua y la Comunicación	http://lct-master.org/
Universidad Autónoma de Barcelona (UAB)	Facultad de Filosofía y Letras	Master TICOM. Tratamiento de la Información y Comunicación Multilingüe	http://www.uab.cat/doc/informacion TICOM
Universidad Autónoma de Madrid		Máster Universitario en Ingeniería Informática y de Telecomunicación	http://www.uam.es/ss/Satellite/es/1242648654244/1242651593660/doctorado/doctorado/Doctorado en Ingeniería Informática y de Telecomunicación.htm
Universidad de Alicante	Escuela Politécnica Superior	Máster Universitario Oficial en Tecnologías de la Informática	http://www.iuii.ua.es/master_ti/
Universidad de Granada		Doble titulación en Traducción y Traductología, y Traducción e Interpretación	https://www.ugr.es/

Tabla 6: Otros productos de PLN y TA en las TIC

Producto	Funcionalidad	NLP, TA	Idiomas
<u>TermiGai</u> <u>UZEI</u>	Identificación y análisis automático de candidatos a término en masas de texto. Utiliza los lematizadores automáticos euLEMA y esLEMA y el extractor de unidades léxicas multipalabra Koloka.	NLP	Vasco Español
<u>euLEMA</u> <u>UZEI</u>	Lematizador automático creado por UZEI para el euskera	NLP	Vasco
<u>esLEMA</u> <u>UZEI</u>	Lematizador automático creado por UZEI para el castellano	NLP	Español
<u>Gaika</u> <u>UZEI</u>	Esta herramienta tiene por objeto determinar la temática de documentos de texto utilizando la categorización documental.	NLP	Vasco-español
<u>Lex2</u> <u>UZEI</u>	Herramienta para la creación semiautomática de diccionarios utiliza corpus paralelos para detectar y extraer automáticamente los pares de términos de las correspondencias en euskera y español	NLP	Vasco-español
<u>iTXEK</u> <u>UZEI</u>	Herramienta para garantizar la calidad y coherencia de los textos que se traducen al euskera, optimizando el proceso de traducción y haciéndolo más rentable.	NLP	Vasco
<u>Hobelex</u> <u>UZEI</u>	Desarrollo de un corrector ortográfico y léxico de Euskera para Microsoft Office	NLP	Vasco
Suite de Herramientas de PLN <u>Elhuyar</u>	– Reconocimiento de personas, lugares, organizaciones y detección de hechos y eventos. – Categorización de temas y extracción de palabras clave. – Propuesta de sinónimos y antónimos. – Corrección ortográfica automática o asistida en cualquier idioma. – Detección de idioma. – Etiquetado Part-Of-Speech y análisis lingüístico. – Extracción de relaciones en el texto y desambiguación.	NLP	– Árabe – Vasco – Catalán – Chino – Inglés – Francés – Alemán – Español – Portugués – Otros
<u>Lucy LT</u>	– Herramienta de traducción automática para facilitar la comunicación a nivel internacional entre diversos sectores empresariales. – Cliente Web y Desktop – LT Task Scheduler (balanceador de carga) – Motores LT (servidores de traducción): nativos o externos, p.e. Moses – LT Lexshop: Gestor de diccionarios para incorporar al sistema su terminología específica.	TA	Traduce entre <u>36 direcciones de traducción</u> y también convierte textos entre diferentes variantes de algunas lenguas.

Producto	Funcionalidad	NLP, TA	Idiomas
Sistema de TA de <u>MondragonLingua</u> con tecnologías Moses SMT	– Motor personalizable de Traducción Automática – Supone una reducción significativa del volumen de palabras a traducir por los traductores humanos, con el consiguiente ahorro en costes. – Genera contenido multilingüe, con el cual MondragonLingua optimizará el servicio de traducción a sus clientes en términos de coste y eficiencia.	TA	
<u>ATLAS Real Time</u> de Linguaserve	– Traducción en tiempo real (Real Time) o de forma muy rápida (Quick Time) –	TA	
<u>Websays Clippings</u>	– Herramienta de análisis y categorización de textos – Resumen y división en categorías de los asuntos principales – Análisis de sentimiento	NLP	
<u>Websays Alerts</u>	– Herramienta que monitoriza campañas – Rastreador de campañas y crisis	NLP	

Tabla 7: Servicios de la red SARA

Aplicación	Descripción
BADARAL	Sistema de gestión descentralizada de Recursos Humanos
NEDAES	Nómina Estándar de la Administración del Estado
RCP	Registro Central de Personal de la AGE
RCPDOC	Sistema de Gestión de Documentación Registral
ADISE	Sistema de Análisis de Disponibilidad de servicios
AFIRMA	Plataforma de Validación de Firma Electrónica
BBDD-060	Sistema de información de legislación, becas y empleo
BUZON-060	Buzón 060
CES	Servicio de Comunicaciones Electrónicas Seguras
CIRCA	Sistema de Gestión Documental para Grupos de Trabajo
CISE	Centro de Interoperabilidad Semántica
CM-OBSAE	Cuadro de Mando del Observatorio de Administración Electrónica
CPCSAE	Gestión de Reuniones de CPCSAE
DIR2	Directorio de unidades administrativas v2
DIR3	Directorio de unidades administrativas v3
DOCUMENTA	Almacén de Documentos de RRHH
EMP-060	Empresas 060
EPAGO	Plataforma de Pago
ESC-060	Escritorio 060
ESIR	Sistema de Información de Recursos Humanos
FAQ-060	FAQs, citas y ticketing
FEP	Servicio Web de Fichas de Empleo Público
FRAMEWORK	Framework de desarrollo de servicios y componentes
FUNCIONA	Portal del Empleado Público
INTERMEDIACIÓN	Plataforma de Intermediación
IPS	Sistema de Inscripción de Pruebas Selectivas
MPAE	Informes OBSAE dispositivos móviles
NOTIFICA	Gestión de notificaciones de RRHH

Aplicación	Descripción
PAE	Portal de Administración Electrónica
PORTAFIRMAS	Portafirmas
PORTAL060	Portal 060
SCCD	Portal de Comunicación de Cambio de Domicilio
PORTALES MMCC	Módulos comunes Portales Magnolia: GUS, CAS, MCC
PORTALFIRMA	Portal de Promoción de Firma Electrónica
REINA	Registro Electrónico de Apoderamientos
REINA	Inventario de Recursos Informáticos de la AGE
SEDE-060	Sede 060
SGC-EELL	SGC-Entes Locales
SIA	Sistema de Información Administrativa
SIR/REC	Sistema de Intercambio de Registros
SIM/SMS	Plataforma de mensajería
SOPORTECAID	Solicitud de soporte técnico al CAID
STORK	Sistema Europeo de Validación de Certificados
VALIDE	Portal de Validación de Certificados y Firma Electrónica
TSA	Plataforma de Sellado de Tiempo
PLATA	Plataforma de Traducción Automática
EUGO	Ventanilla Única de la Directiva de Servicios
TECNIMAP	Web de información de antiguos Tecnimaps
FORJA-CTT	Forja del Centro de Transferencia de Tecnología
MARCA-ESPAÑA	Web de Marca España
FORMA	Formularios Electrónicos
FACE	Factura Electrónica y Certificado de Administración
SIGP	Sistema Integrado de Gestión de Personal
TRAMA	Frontal DMZ de tramitación de Asuntos Laborales
FICHAJES	Frontal DMZ del sistema de Control de Fichajes
EGEO	Sistema de Georreferenciación
AUTENTIC@	Sistema de Autenticación Centralizado
AVISAME	Servicio de gestión de notificaciones

Aplicación	Descripción
CLAVE	Intermediador de Sistemas de Autenticación
RIS	Registro Integrado Sectorial
TRANSPARENCIA	Portal de implementación Ley de Transparencia
MERCADO	Portal de implementación Ley de Unidad de Mercado

9.7. Tabla 8: Contratos en tecnologías del lenguaje de la Plataforma de Contratación del Estado

TIPO	Fecha	Precio	Contratante	Objeto del Contrato
Automatic Call Distributor (ACD)	19/06/15	2.664.088 €	Red.es	Servicios de Atención y Soporte a Usuarios, Exp.: 095/15-SI -Sistema avanzado de enrutamiento de llamadas
Buscador y semántica	18/06/15	3.500.000 €	Patrimonio Nacional	Servicio de Implantación de la Solución de Preservación Digital, y del Metabusador para las Colecciones Reales de Patrimonio Nacional
Buscador	12/03/15	200.000 €	ICO	Solución de Gestión Documental Corporativa para el Instituto de Crédito Oficial- Buscador Textual
Análisis noticias	25/07/14	266.607 €	EOI	Plataforma Tecnológica para Generar un Cuadro de Mando en los Destinos Turísticos Inteligentes - Procesamiento noticias
	22/07/13	3.267.798 €	Red.es	Servicio de Atención y Soporte para la Oficina de Atención al Usuario de Telecomunicaciones - 053/13-SI - Automatic Call Distributor (ACD)
Automatic Call Distributor (ACD)	18/07/13	2.273.968 €	Correos	Atención Telefónica a Clientes y Usuarios de Correos - Automatic Call Distributor (ACD)
Buscador y semántica	05/07/13	446.372 €	Red.es	Sistema de Gestión de Archivos Judiciales y Ubicaciones (Agu) en el Ámbito de la Administración de Justicia de Cataluña - Buscador y semántica
Automatic Call Distributor (ACD) - castellano y lenguas cooficiales	01/10/12	607.685 €	Red.es	Servicios de Atención y Soporte a Usuarios Procedimiento Abierto - 085/12-SI
Automatic Call Distributor (ACD)	21/05/12	1.211.813 €	Red.es	Servicio de Atención al Ciudadano para la Oficina de Atención al Usuario de las Telecomunicaciones y la TDT
Buscador	30/09/11	177.000 €	Fremap	Plataforma de Gestión de la Formación "On Line" y Presencial de Accidentes de Trabajo y Enfermedades Profesionales
Buscador	28/07/11	1.292.992 €	Correos	Web publica de correos
Buscador	22/07/11	177.000 €	Fremap	Plataforma de gestión de la formación "on line" y presencial de accidentes de trabajo y enfermedades profesionales de la FREMAP, mutua Seguridad Social
Automatic Call Distributor (ACD)	28/02/11	5.044.040 €	Correos	Servicio de atención telefónica a clientes y usuarios de correos
Traducción automática	18/12/14	17.999 €	AEAT	Servicio de traducción automática externo para la presentación en diferentes lenguas cooficiales (catalán, gallego y valenciano) y en idioma inglés de

TIPO	Fecha	Precio	Contratante	Objeto del Contrato
Traducción automática	27/11/14	16.965 €	MINHAP Dirección de Tecnologías de la Información y las Comunicaciones	las páginas Web del portal de la Agencia Tributaria (excluida la Sede electrónica) durante los meses de enero y febrero de 2015. Servicios para la actualización y soporte de la plataforma de traducción automática Plata
Traducción automática	18/03/13	54.272 €	INTECO	Suministro, instalación y mantenimiento de servidores para la Plataforma de Traducción Automática del Ministerio de Hacienda y Administraciones Públicas
Traducción automática	24/07/12	120.786 €	INTECO	Puesta en marcha de una Plataforma de traducción Automática.
Traducción automática	15/09/10	272.101 €	AEAT	Contratación de un servicio de traducción automática externo para la presentación en diferentes lenguas cooficiales (catalán, gallego y valenciano) de las páginas Web del portal de la Agencia Tributaria (excluida la sede)
Traducción automática	15/09/10	288.136 €	AEAT	Servicios de traducción a las lenguas cooficiales y otros idiomas extranjeros de los contenidos ofrecidos desde los portales en Internet de la Agencia Tributaria y del contenido de comunicaciones escritas dirigidas a los contribuyentes con destino al Departamento de Informática Tributaria de la AEAT, durante dos años y desglosado en dos lotes
Traducción automática	31/07/10	127.500 €	Presidencia Española del Consejo de la Unión Europea	Contratación de un Servicio De Traducción de los Contenidos de la página web de la Presidencia Española del Consejo de la Unión Europea del Primer Semestre de 2010