|  |
| --- |
|  |
|  |
| Serie de Libros Blancos de META-NET |
| Las lenguas en la  Sociedad de la Información Europea |
| – Español – |
|  |
|  |

|  |  |  |
| --- | --- | --- |
| flag_2colors.jpg | FP7-gen-CMYK.jpg | logo_ict_psp.2.gif |

La publicación de este documento ha sido financiada por el Séptimo Programa Marco y el Programa de Apoyo a Políticas TIC de la Comisión Europea en virtud de los contratos T4ME (Acuerdo de subvención del 249 119), CESAR (Acuerdo de subvención del 271 022), METANET4U (Acuerdo de Subvención 270.893) y Meta-NORD (Acuerdo de subvención del 270 899).

# Prólogo

Este documento es parte de una serie de Libros Blancos que promueve el conocimiento sobre las tecnologías del lenguaje y sobre su potencial. Se dirige a educadores, periodistas, políticos y las comunidades lingüísticas, entre otros.

La disponibilidad y el uso de la tecnología lingüística en Europa varían entre los distintos idiomas. En consecuencia, las acciones que se requieren para apoyar la investigación y el desarrollo de las tecnologías de la lengua también son diferentes para cada idioma. Las acciones necesarias dependen de muchos factores, tales como la complejidad de un lenguaje determinado y el tamaño de su comunidad.

META-NET, una red de excelencia financiada por la Comisión Europea, ha llevado a cabo un análisis de los recursos lingüísticos y las tecnologías actuales. Este análisis se ha centrado en las 23 lenguas oficiales europeas, así como en otros idiomas importantes, a nivel nacional y regional en Europa. Los resultados de este análisis sugieren que hay muchas lagunas en la investigación relevante para cada idioma. Un análisis más detallado por parte de los expertos y una evaluación de la situación actual contribuirá a maximizar el impacto de las nuevas investigaciones y a minimizar los riesgos.

META-NET se compone de 47 centros de investigación de 31 países que están trabajando con representantes de empresas comerciales, agencias gubernamentales, industria, organizaciones de investigación, empresas de software, proveedores de tecnología y universidades europeas. Juntos están creando una visión tecnológica común, y al mismo tiempo están desarrollando una agenda estratégica de investigación que, a 10 años vista, permita a las aplicaciones basadas en tecnología lingüística abordar las deficiencias detectadas.

Contacto

META-NET

DFKI Projektbüro Berlin

Alt-Moabit 91c

10559 Berlin

Germany

office@meta-net.eu

http://www.meta-net.eu

Autor

Dr. Maite Melero, Barcelona Media Centre d‘Innovació

Table of Contents

Prólogo 3

Resumen ejecutivo 6

Un riesgo para nuestras lenguas y un reto para la tecnología lingüística 9

Las fronteras lingüísticas son un obstáculo para la Sociedad de la Información Europea 10

Nuestras lenguas en peligro 10

La tecnología lingüística es clave en el desarrollo de muchas aplicaciones 11

Oportunidades para la tecnología lingüística 11

Retos de la tecnología lingüística 12

Adquisición del lenguaje por los seres humanos y por las máquinas 13

El español en la Sociedad de la Información Europea 15

Datos generales 15

Particularidades de la Lengua Española 15

Acontecimientos recientes 16

El cultivo del idioma en España 17

La lengua en la educación 19

El español en Internet 21

Recursos de tecnología lingüística para el español 24

Arquitectura de las aplicaciones 24

Áreas principales de aplicación 25

Corrección de textos 25

Búsquedas en la Web 26

Procesamiento del habla 28

Traducción Automática 29

Otras áreas de aplicación 32

Programas Educativos 34

Proyectos e iniciativas nacionales 34

Disponibilidad de herramientas y recursos 35

Comparación entre lenguas 37

Conclusiones 39

Acerca de META-NET 41

Lineas de Acción 42

Organizaciones miembros 43

Referencias 46

# Resumen ejecutivo

**La tecnología lingüística construye puentes para el futuro de Europa**

Durante los últimos 60 años, Europa ha adquirido una estructura política y económica singular y, sin embargo, cultural y lingüísticamente todavía es muy diversa. Esto implica que, del portugués al polaco y del italiano al islandés, la comunicación cotidiana entre los ciudadanos europeos, así como la comunicación en el ámbito de los negocios y de la política se enfrenta inevitablemente con las barreras del idioma. Las instituciones de la UE gastan casi mil millones de euros al año en mantener su política de multilingüismo, es decir, traduciendo documentos e interpretando intervenciones orales. Sin embargo, ¿es necesario que estas tareas sigan siendo una carga pesada y costosa? Las modernas tecnologías de la lengua y la investigación lingüística pueden contribuir significativamente a derribar las barreras lingüísticas. En el futuro, las tecnologías lingüísticas, combinadas con dispositivos y aplicaciones inteligentes, serán capaces de ayudar a los europeos a comunicarse fácilmente entre sí y a hacer negocios, incluso si no hablan una lengua común.

El mercado único europeo resulta beneficioso para los países que lo integran. Sin embargo, las barreras del idioma pueden suponer un freno para los negocios, especialmente para las pequeñas y medianas empresas, que no tienen los medios económicos para afrontar la situación. La única (e impensable) alternativa a esta Europa multilingüe sería permitir que una sola lengua tomara una posición dominante y terminara reemplazando al resto de lenguas.

El aprendizaje de lenguas extranjeras siempre ha sido una forma habitual de superar las barreras lingüísticas. Sin embargo, sin apoyo tecnológico, llegar a dominar las 23 lenguas oficiales de los Estados miembros de la Unión Europea, sumadas a 60 lenguas más no oficiales, supone un obstáculo insuperable para los ciudadanos de Europa así como para su economía, para el debate político y para el progreso científico.

La solución consiste en invertir en las tecnologías clave para la superación de estas barreras, es decir, las tecnologías lingüísticas. Estas tecnologías ofrecen ventajas enormes, no sólo dentro del mercado común europeo, sino también en las relaciones comerciales con terceros países, especialmente en las economías emergentes. Para lograr este objetivo, y preservar la diversidad cultural y lingüística de Europa, es conveniente considerar las particularidades lingüísticas de todos los idiomas europeos, y analizar el estado actual de las tecnologías lingüísticas para cada uno de ellos. Las tecnologías lingüísticas constituirán en el futuro un puente único entre las lenguas de Europa.

**Las tecnologías de la lengua, clave para el futuro**

La traducción automática y las herramientas de procesamiento del habla que están actualmente disponibles en el mercado, aún están lejos de alcanzar este ambicioso objetivo. Los agentes dominantes en estos ámbitos son principalmente empresas de propiedad privada radicadas en América del Norte. Ya en la década de 1970, la UE se dio cuenta de la enorme relevancia de la tecnología lingüística como conductor de la unidad europea, y comenzó a financiar sus primeros grandes proyectos de investigación, como EUROTRA. Al mismo tiempo, se establecieron proyectos nacionales que generaron resultados valiosos, pero que nunca llevaron a una acción europea concertada. En contraste con este esfuerzo financiero altamente selectivo, otras sociedades multilingües, como la India (22 lenguas oficiales) y Sudáfrica (11 lenguas oficiales) han creado recientemente programas nacionales a largo plazo para la investigación lingüística y el desarrollo tecnológico.

Muchas de las aplicaciones de tecnología lingüística utilizan actualmente métodos estadísticos, que se basan en grandes cantidades de datos e ignoran las propiedades intrínsecas de la lengua. Por ejemplo, algunos de los sistemas de traducción más populares traducen automáticamente mediante la comparación de la frase a traducir con centenares de miles de frases previamente traducidas por humanos. La calidad de la producción depende en gran medida de la cantidad y la calidad de la muestra de corpus disponible. Así, mientras que la traducción automática de oraciones sencillas, en idiomas con una cantidad suficiente de texto disponible, puede obtener resultados útiles, los métodos estadísticos puros están condenados al fracaso en el caso de idiomas con cantidades mucho menores de datos, o en el caso de las construcciones sintácticas complejas. Dada esta situación, la Unión Europea ha decidido financiar proyectos tales como EuroMatrix y EuroMatrixPlus (desde 2006) y iTranslate4 (desde 2010) que llevan a cabo investigación básica y aplicada y generan recursos lingüísticos para todos los idiomas europeos. El análisis de las propiedades estructurales más profundas de la lengua es el único camino posible si queremos crear aplicaciones que funcionen bien para toda la gama de las lenguas de Europa.

La investigación europea en este ámbito ya ha logrado varios éxitos. Por ejemplo, los servicios de traducción de la Unión Europea actualmente están utilizando MOSES, una aplicación de traducción automática de código abierto, que se ha desarrollado principalmente a través de proyectos de investigación europeos. Muchos de los laboratorios de investigación y desarrollo (por ejemplo, IBM y Philips) han cerrado o se han trasladado a otro lugar. En lugar de construir sobre los resultados de sus proyectos de investigación, Europa ha tendido a realizar actividades de investigación aisladas, con menor impacto en el mercado. Incluso el valor económico de estos primeros esfuerzos, puede verse en el número de spin-offs. Una compañía como Trados, que fue fundada en 1984, fue vendida a SDL, con sede en el Reino Unido, en 2005.

**Las tecnologías lingüísticas contribuyen a unificar Europa**

Basándose en los conocimientos acumulados hasta el momento, parece claro que la actual tendencia a la tecnología híbrida, mezcla de procesamiento lingüístico de la lengua con métodos estadísticos, será capaz de reducir la brecha entre todas las lenguas europeas e ir más allá. Como esta serie de “libros blancos” muestra, existen grandes diferencias de preparación entre los diferentes idiomas y estados europeos con respecto a las tecnologías de la lengua. Sin embargo, incluso los idiomas “grandes” como el español, todavía necesitan dedicar más recursos a la investigación con objeto de que las soluciones tecnológicas estén realmente listas para su uso cotidiano.

El objetivo a largo plazo de META-NET es introducir tecnología lingüística de alta calidad para todas las lenguas a fin de lograr la unidad política y económica a través de la diversidad cultural. La tecnología ayudará a derribar las barreras existentes y a construir puentes entre las lenguas de Europa. Esto requiere que todas las partes interesadas – del mundo de la política, de la investigación, la industria y la sociedad - unan sus esfuerzos de cara al futuro.

Esta serie de “libros blancos” complementa otras acciones estratégicas adoptadas por META-NET (véase el apéndice para un resumen). En el sitio web de META-NET (<http://www.meta-net.eu>) puede hallarse información actualizada, como la versión actual del documento de visión META-NET, o la Agenda Estratégica de Investigación (Strategic Research Agenda o SRA).

# Un riesgo para nuestras lenguas y un reto para la tecnología lingüística

Somos testigos de una revolución digital comparable a la invención de la imprenta por Gutenberg.

Somos testigos de una revolución digital que está impactando drásticamente en la comunicación y en la sociedad. Esta revolución se ha comparado alguna vez con la invención de la imprenta por Gutenberg. ¿Qué nos puede decir esta analogía sobre el futuro de la sociedad de la información y de nuestras lenguas en particular?

Después de la invención de Gutenberg, los avances reales en la comunicación y el intercambio de conocimientos se lograron mediante esfuerzos tales como la de la Biblia por Lutero a la lengua vernácula. En los siglos siguientes, se desarrollaron técnicas culturales para manejar mejor el procesamiento del lenguaje y el intercambio de conocimiento:

* la estandarización ortográfica y gramatical de las principales lenguas permitió la rápida difusión de nuevas ideas científicas e intelectuales;
* el desarrollo de las lenguas oficiales hizo posible que los ciudadanos pudieran comunicarse dentro de ciertos límites (a menudo políticos);
* la enseñanza de lenguas y las traducciones favorecieron los intercambios entre las distintas lenguas y culturas;
* la creación de editoriales y de directrices bibliográficas aseguraron la calidad y disponibilidad de material impreso;
* la creación de distintos medios de comunicación como periódicos, radio, televisión, libros y otros formatos satisfizo las diferentes necesidades de comunicación.

En los últimos veinte años, la tecnología de la información ha ayudado a automatizar y facilitar muchos de los procesos:

* el software de autoedición ha sustituido la mecanografía y la composición;
* Microsoft PowerPoint ha sustituido las transparencias para retroproyector;
* el correo electrónico envía y recibe documentos más rápido que una máquina de fax;
* Skype ofrece llamadas baratas de telefonía por Internet y organiza reuniones virtuales;
* los formatos de audio y de codificación de vídeo facilitan el intercambio de contenido multimedia;
* los motores de búsqueda basada en palabras clave proporcionan acceso a las páginas web;
* servicios en línea como Google Translate producen traducciones rudimentarias de forma rápida;
* las plataformas de medios sociales como Facebook, Twitter y Google+ facilitan la comunicación, la colaboración y el intercambio de información.;

A pesar de que tales herramientas y aplicaciones son útiles, no son todavía capaces de dar soporte a una sociedad europea multilingüe y sostenible, donde la información y las mercancías puedan circular libremente.

## Las fronteras lingüísticas son un obstáculo para la Sociedad de la Información Europea

Un espacio económico y de información globalizado supone una mayor cantidad de lenguas, hablantes y contenido.

No podemos predecir exactamente como será la sociedad de la información del futuro, pero muy probablemente la gran revolución en la tecnología de la comunicación será relacionar de maneras innovadoras a las personas que hablan diferentes idiomas. Más idiomas y más contenidos interactuarán con nuevos medios de comunicación en un mismo espacio económico y de información globalizado. La actual popularidad de las redes sociales (Wikipedia, Facebook, Twitter, YouTube, y, recientemente, Google +) es sólo la punta del iceberg.

Hoy en día, podemos transmitir gigabytes de texto alrededor del mundo en pocos segundos, incluso antes de reconocer que el mensaje es en un idioma que no entendemos. Según un reciente informe de la Comisión Europea, el 57% de los usuarios europeos de Internet adquieren bienes y servicios en un idioma distinto de su lengua materna. (El idioma más común es el inglés, seguido del francés, el alemán y el español.) El 55% de los usuarios leen contenidos en un idioma extranjero, mientras que sólo el 35% usa otra lengua para escribir comentarios, correos electrónicos o publicar contenido en la Web.[[1]](#endnote-1) Hace algunos años el inglés era la lengua franca de la Web -la gran mayoría de los contenidos estaba en inglés- pero la situación ha cambiado drásticamente. La cantidad de contenido en línea en otras lenguas europeas (así como asiáticas) se ha disparado.

Curiosamente, esta brecha digital causada por las barreras lingüísticas no recibe mucha atención en los medios y, sin embargo, la cuestión que se plantea es crucial: ¿Qué idiomas europeos prosperarán en la red y en la sociedad digital, y cuáles están condenados a desaparecer?

## Nuestras lenguas en peligro

Así como la imprenta ayudó a intensificar el intercambio de información en Europa, también llevó a la extinción de muchas lenguas europeas. Ciertas lenguas regionales y minoritarias, como el córnico y el dálmata, tuvieron una producción impresa escasa y su transmisión se limitó a las formas orales, lo que a su vez limitó su ámbito de uso. ¿Tendrá Internet el mismo impacto en nuestras lenguas?

Los aproximadamente 60 idiomas de Europa son uno de sus bienes culturales más ricos e importantes, y una parte vital de su singular modelo social. [[2]](#endnote-2) Si bien algunos de sus idiomas, como el inglés y el español, probablemente sobrevivan en el mercado digital emergente, muchos otros podrían convertirse en irrelevantes en una sociedad en red. Esto debilitaría la posición global de Europa, e iría en contra del objetivo estratégico de garantizar la participación equitativa de todos los ciudadanos europeos, independientemente de su idioma. De acuerdo con un informe de la UNESCO en materia de multilingüismo, las lenguas son un medio esencial para el disfrute de derechos fundamentales como la expresión política, la educación y la participación en la sociedad.[[3]](#endnote-3)

## La tecnología lingüística es clave en el desarrollo de muchas aplicaciones

En el pasado reciente, la mayoría de la inversión destinada a la preservación de las lenguas se centraban en la enseñanza de idiomas y en la traducción. Según una estimación, el mercado europeo de traducción, interpretación, localización de software y globalización de sitios web fue de 8.4 mil millones de euros en 2008 y se espera que crezca en un 10% anual. [[4]](#endnote-4) Sin embargo, esta cifra sólo cubre una pequeña proporción de las necesidades actuales y futuras de la comunicación entre lenguas. La solución más efectiva para garantizar el multilingüismo en la Europa del futuro pasa por el uso de tecnologías apropiadas, de la misma forma que recurrimos a la tecnología para resolver nuestras necesidades de transporte, de energía y de atención a la discapacidad, entre otras.

La tecnología lingüística (orientada a todas las formas de texto escrito y discurso hablado) ayuda a las personas a colaborar entre ellas, a hacer negocios, a compartir conocimientos y a participar en los debates sociales y políticos, independientemente de las barreras idiomáticas y de las habilidades informáticas. A menudo, opera de forma invisible dentro de sistemas de software complejos para ayudarnos a:

La tecnología lingüística ayuda a las personas a colaborar entre ellas, a hacer negocios, a compartir conocimientos y a participar en debates sociales y políticos, por encima de las barreras lingúísticas.

* encontrar información con un motor de búsqueda por Internet;
* comprobar la ortografía y la gramática en un procesador de textos;
* ver recomendaciones de productos en una aplicación de comercio electrónico;
* escuchar las instrucciones verbales de un sistema de navegación para automóviles;
* traducir páginas web a través de un servicio en línea.

Las tecnologías de la lengua consisten en una serie de aplicaciones básicas que operan en un marco de aplicación más amplio. El propósito de la serie de White Papers de META-NET es describir el grado de preparación de cada una de las lenguas europeas en estas tecnologías básicas..

Para mantener nuestra posición en la vanguardia de la innovación global, Europa necesitará disponer de tecnologías lingüísticas adaptadas a todos los idiomas europeos, que sean sólidas, asequibles y estén perfectamente integradas en los entornos de software clave. En un futuro cercano, sin la tecnología lingüística, no seremos capaces de lograr una experiencia de usuario interactiva, multimedia y multilingüe realmente efectiva.

Europa necesita tecnología lingüística robusta y asequible para todos los idiomas europeos.

## Oportunidades para la tecnología lingüística

En el mundo de la imprenta, el gran avance tecnológico fue la duplicación rápida de la imagen de un texto (una página), utilizando una prensa. Los seres humanos aún tenían que hacer el trabajo laborioso de buscar, leer, traducir y resumir el conocimiento. Tuvimos que esperar hasta Edison para grabar el lenguaje hablado, y en este caso también su tecnología simplemente hacía copias analógicas.

Las tecnologías lingüísticas son capaces de automatizar los procesos mismos de traducción, producción de contenidos y gestión del conocimiento para todos los idiomas europeos. También pueden servir para construir interfaces intuitivas, basadas en lenguaje humano (escrito o hablado) para electrodomésticos, maquinaria, vehículos, computadoras y robots. Las aplicaciones comerciales e industriales de estas tecnologías se encuentran todavía en las primeras etapas de desarrollo, sin embargo, los éxitos de la I + D están creando una ventana real de oportunidades. Por ejemplo, la traducción automática ya es razonablemente precisa en áreas concretas, y existen aplicaciones experimentales que proporcionan información multilingüe y gestión del conocimiento, así como producción de contenidos, en muchas lenguas europeas.

Como ocurre con la mayoría de las tecnologías, las primeras aplicaciones lingüísticas, como las interfaces basadas en voz y los sistemas de diálogo, se desarrollaron para ámbitos altamente especializados, y con frecuencia su rendimiento es limitado. Sin embargo, existen grandes oportunidades de mercado en ámbitos como la educación o el ocio, con la integración de tecnología lingüística en juegos, en divulgación del patrimonio cultural, en paquetes de entretenimiento educativo, en bibliotecas, entornos de simulación y programas de capacitación. Los servicios de información móvil, el software de aprendizaje de idiomas, los entornos de e-learning, las herramientas de autoevaluación y el software de detección de plagio son sólo algunas de las áreas de aplicación en las que la tecnología lingüística puede desempeñar un papel importante. La popularidad de las redes sociales como Twitter y Facebook sugieren necesidades adicionales para esta tecnología, como por ejemplo, control de foros, resumen de discusiones, tendencias de opinión, detección de respuestas emocionales o identificación de infracciones de derechos de autor, entre otros.

La tecnología lingüística representa una gran oportunidad para la Unión Europea y puede ayudar a abordar el complejo tema del multilingüismo en Europa - el hecho de que distintas lenguas convivan de forma natural en empresas, organizaciones y escuelas europeas. Los ciudadanos tienen que comunicarse a través de estas fronteras lingüísticas que cruzan el Mercado Común Europeo, y la tecnología lingüística puede ayudar a superar esta barrera final, a la vez que apoya el uso libre y abierto de cada una de las lenguas individuales. Mirando aún más hacia adelante, una tecnología lingüística multilingüe e innovadora, proporcionará un punto de referencia a nuestros socios a nivel mundial, cuando empiecen a atender a sus propias comunidades multilingües. La tecnología lingüística puede verse como una forma de tecnología “asistencial” que ayuda a superar la "discapacidad" de la diversidad lingüística y hace que las comunidades lingüísticas sean más accesibles entre sí.

La tecnología lingüística ayuda a superar la "discapacidad" de la diversidad lingüística.

Por último, un campo activo de investigación es el innovador uso de la tecnología lingüística en operaciones de rescate en zonas de desastre, donde la eficiencia puede ser una cuestión de vida o muerte: los robots inteligentes del futuro, dotados de capacidades multilingües tendrán un potencial aumentado para salvar vidas.

## Retos de la tecnología lingüística

Aunque la tecnología lingüística ha hecho progresos considerables en los últimos años, el ritmo actual de progreso tecnológico y la innovación en los productos es demasiado lento. Incluso tecnologías ampliamente utilizadas, como la corrección ortográfica y gramatical de los procesadores de texto, sólo están disponibles para unos cuantos idiomas. Los servicios de traducción automática en línea, aunque son útiles para generar de forma rápida una aproximación razonable al contenido de un documento, todavía no resultan prácticos cuando la traducción debe ser muy exacta. Debido a la complejidad del lenguaje humano, el desarrollo del software que lo modela y su puesta a prueba en el mundo real resulta un proceso largo y costoso que requiere de financiación continuada. Europa debe pues mantener su papel pionero a la hora de enfrentarse a los desafíos tecnológicos de una comunidad multilingüe inventando nuevos métodos para acelerar el desarrollo transversal de esta. Estos nuevos métodos incluyen tanto avances computacionales como técnicas emergentes, como por ejemplo, el crowdsourcing.

El ritmo actual de progreso tecnológico y la innovación en los productos es demasiado lento.

## Adquisición del lenguaje por los seres humanos y por las máquinas

Para ilustrar cómo manejan el lenguaje los ordenadores y por qué es difícil programarlos para usarlo, vamos a examinar brevemente la manera en que los seres humanos adquieren sus primeras y segundas lenguas, y luego cómo funcionan las tecnologías lingüísticas.

Los seres humanos adquieren capacidades lingüísticas de dos maneras diferentes. Los bebés aprenden un idioma al escuchar las interacciones reales entre sus padres, hermanos y personas del ámbito familiar. Desde los dos años, los niños producen sus primeras palabras y frases cortas. Esto sólo es posible porque los seres humanos tienen una predisposición genética a imitar y racionalizar lo que escuchan.

Los seres humanos adquieren capacidades lingüísticas de dos maneras diferentes: aprendiendo de los ejemplos y aprendiendo las reglas gramaticales subyacentes.

Aprender un segundo idioma a una edad avanzada requiere más esfuerzo, sobre todo si la persona no se encuentra inmersa en una comunidad lingüística de hablantes nativos. En la escuela, los idiomas extranjeros se adquieren a través del aprendizaje explícito del vocabulario y la gramática, haciendo ejercicios, aplicando reglas abstractas y viendo ejemplos. El aprendizaje de una lengua extranjera se hace más difícil con la edad.

Existen un paralelismo claro entre estas dos maneras de aprender lenguas en los seres humanos y los dos tipos principales de sistemas de tecnología lingüística: los sistemas estadísticos o basados en los datos, y los sistemas basados en reglas.

Existen un paralelismo claro entre las dos maneras de aprender lenguas en los seres humanos y los dos tipos principales de sistemas de tecnología lingüística: los estadísticos y los basados en reglas.

Los sistemas estadísticos obtienen conocimiento lingüístico a partir de grandes colecciones de textos concretos. El algoritmo de aprendizaje automático "aprende" cómo se traducen las palabras, las frases cortas y, por último, las oraciones completas. Los textos no son siempre fáciles de obtener. Si bien para entrenar ciertas aplicaciones, por ejemplo, un corrector ortográfico, es suficiente usar texto en un solo idioma, para entrenar un sistema de traducción automática se precisan textos paralelos, en dos o más lenguas. El enfoque estadístico puede requerir millones de frases y su rendimiento aumenta con la cantidad de texto analizado. Esta es una de las razones por la que los proveedores de motores de búsqueda están dispuestos a recoger tanto material escrito como sea posible. Servicios como Google Search y Google Translate se basan en métodos estadísticos. La gran ventaja de la estadística es que la máquina aprende rápidamente en una serie sucesiva de ciclos de entrenamiento, a pesar de que la calidad del resultado puede variar de forma aparentemente arbitraria.

La estrategia alternativa es la construcción de sistemas basados ​​en reglas. Expertos lingüistas e informáticos primero tienen que codificar el conocimiento gramatical o las reglas de traducción y compilar diccionarios y listas de vocabulario. Esto requiere una gran inversión en tiempo y esfuerzo. Algunos de los principales sistemas de traducción automática basados ​​en reglas han sido objeto de constante desarrollo durante más de veinte años. La gran ventaja de los sistemas basados ​​en reglas es que los expertos tienen un control directo sobre el procesamiento del lenguaje. Esto hace que sea posible corregir errores sistemáticos y dar información detallada al usuario, especialmente cuando los sistemas de reglas se utilizan para el aprendizaje de idiomas. Sin embargo, debido a su alto coste, este tipo de sistemas sólo han sido desarrollados para los idiomas mayoritarios.

Como las ventajas e inconvenientes de los sistemas estadísticos y de los basados en reglas tienden a ser complementarios, la investigación actual se centra en los enfoques híbridos, que combinan las dos metodologías. Sin embargo, estos métodos de momento han tenido más éxito en los centros de investigación que en las aplicaciones industriales.

Como hemos visto en este capítulo, muchas aplicaciones de las más comunes dependen en gran medida la tecnología lingüística. Por su condición multilingüe, esto es particularmente crítico en el espacio europeo de la información. Aunque la tecnología lingüística ha hecho considerables progresos en los últimos años, todavía hay un enorme potencial para mejorar su calidad. En las siguientes secciones, vamos a describir el papel del español en la sociedad europea de la información y evaluar el estado actual de la tecnología lingüística para el español.

# El español en la Sociedad de la Información Europea

## Datos generales

El español es hoy uno de los idiomas con más peso demográfico en el mundo. Según Ethnologue**[[5]](#endnote-5)** es el segundo idioma más hablado en términos de hablantes nativos después del chino. Alrededor del 6% de la población mundial, o sea más de 400 millones de hablantes, tienen el español como primera lengua.

El español es hoy uno de los idiomas con más peso demográfico en el mundo.

El español es la lengua oficial de España, donde se originó como una evolución del latín vulgar, pero la mayoría de los hablantes del español se encuentran en América Latina. De todos los países con mayoría de hablantes de español, sólo España y Guinea Ecuatorial se encuentran fuera de América. México tiene el mayor número de hablantes nativos. A nivel nacional, el español es el idioma oficial de Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, República Dominicana, Ecuador, El Salvador, Guatemala, Honduras, México, Nicaragua, Panamá, Paraguay, Perú, Puerto Rico, Uruguay y Venezuela[[6]](#endnote-6).

En América, el español coexiste con una multitud de lenguas indígenas, algunas de ellas también co-oficiales con el español en sus respectivos países, como el quechua y el aymara en Perú y Bolivia, y el guaraní en Paraguay. En general, la situación se puede caracterizar como diglósica, siendo el español el idioma de prestigio social, mientras que el otro idioma se utiliza como vehículo de comunicación en el ámbito familiar. El español funciona efectivamente como una *koiné* y se caracteriza por su homogeneidad lingüística, especialmente entre los hablantes educados, sin olvidar que cuenta con diversas variedades sociolingüísticas y geolingüísticas.

En España y en algunas partes del mundo de habla española, pero no en todas, el español recibe el nombre de *castellano*, además de español, es decir, el idioma de la región de Castilla, en contraste con otros idiomas que se hablan en España, y que son co-oficiales en sus territorios, como el catalán, el gallego y el vasco.

A pesar de su homogeneidad en general, existen variaciones, principalmente entre la lengua hablada en las zonas centro y norte de España, y la que se habla en gran parte del sur de España, las Islas Canarias y América Latina. Aparte de las variaciones léxicas locales, es bien conocida la diferente pronunciación de la letra c ante las vocales *e* y *i* (como en la palabra *Barcelona*), siendo una fricativa sorda dental / θ / en la primera zona y una sibilante / s / en la segunda. Otras variaciones afectan al uso de los pronombres de segunda persona: *tú* o *vos*, según las zonas, para el singular, y *ustedes* o *vosotros*, para el plural.

## Particularidades de la Lengua Española[[7]](#endnote-7)

Algunas de las características propias del español resultan un reto para el procesamiento computacional.

La fonología española presenta un sistema de cinco vocales. El español es una lengua flexiva, con un sistema de dos géneros y alrededor de cincuenta formas verbales conjugadas, pero flexión limitada de sustantivos, adjetivos y determinantes. Utiliza preposiciones y, por lo general, aunque no siempre, los adjetivos se colocan detrás de los sustantivos, como en la mayoría de las otras lenguas románicas. Su sintaxis es generalmente sujeto-verbo-objeto, aunque las variaciones son frecuentes.

La sintaxis del español marca de forma explícita los objetos directos humanos, que van precedidos por la preposición "a":

*Vi a María* pero *Vi la película*

Puede usarse un pronombre de objeto indirecto (*le*, *les*) “redundante“ incluso en presencia de una frase nominal de objeto indirecto explícita:

*Le di las llaves a María.*

En lo que respecta a los pronombres de sujeto, el español, como la mayoría de lenguas románicas, permite que no estén presentes de froma explícita, de manera que el verbo flexionado puede aparecer solo. En comparación con otras lenguas romances, el español tiene una sintaxis un poco más libre, con menos restricciones en el orden de las palabras.

Debido al prolongado contacto con otras lenguas, el léxico español contiene préstamos del vasco/ibero, germánico, árabe y de lenguas indígenas americanas, aunque alrededor de un 94% de su vocabulario proviene íntegramente del latín.

La ortografía española es mucho más transparente que la ortografía del inglés, por ejemplo. La pronunciación de una palabra puede deducirse fácilmente a partir de su forma escrita. No obstante, algunas de las reglas del sistema de escritura conducen potencialmente a la homofonía. Estas incluyen la <*h*> muda o (*a*/ *ha*), la falta de distinción entre <*b*> y <*v*> (*baca* / *vaca*), o <*g*> y <*j*> antes de <*e*> o <i> (*deje* / *general*), así como algunas fusiones dialectales, como la existente entre <*y*> y <*ll*> (*pollo* / *poyo*), y entre <*c*> o <*z*> y <*s*> (*tensión* / *atención*). Así, la misma pronunciación puede escribirse de diversas maneras. Esta es una fuente de posibles errores que los correctores ortográficos y gramaticales deben ser capaces de detectar y corregir.

La homofonía es una fuente de posibles errores que los correctores ortográficos y gramaticales deben ser capaces de detectar y corregir.

El español escrito señala de manera inequívoca el acento fonético a través de una serie de reglas ortográficas. El acento por defecto está en la penúltima sílaba de las palabras que terminan en vocal, <*n*> o <*s*> y en la última sílaba, cuando la palabra termina en cualquier consonante que no sea <*n*> o <*s*>. Las palabras que no se acentúan por defecto llevan un acento gráfico (agudo) sobre la vocal acentuada.

En español las mayúsculas se usan relativamente poco, comparado con otras lenguas. En general, sólo se ponen en mayúsculas los nombres de personas, de lugar, de empresas u organismos, algunas abreviaturas (por ejemplo, *Sr*., *Sra*.,...), la primera palabra en el título de un libro, película, canción, etc. y la primera palabra de una frase. En cambio se dejan en minúscula los nombres de las nacionalidades y de los idiomas, así como los días de la semana o los meses del año.

## Acontecimientos recientes

El idioma español está creciendo en importancia económica en América Latina y en los EE.UU.

Como se ha dicho, debido a razones demográficas, especialmente fuera de España (en Latinoamérica), el español está escalando posiciones en las filas de los idiomas más hablados y utilizados del mundo.

El idioma está creciendo en importancia económica en América Latina y en los EE.UU., donde compite con el inglés en una serie de poblaciones de habla predominantemente española.

España se promociona activamente como centro mundial de autoridad de la lengua.

España se promociona activamente como centro mundial de autoridad de la lengua, con instituciones de financiación pública, como la Real Academia Española y el Instituto Cervantes y fundaciones privadas, como la Fundéu, que vigila los nuevos usos de la lengua así como la influencia de otras lenguas (ver la siguiente sección).

España también ha hecho progresos ejemplares en el desarrollo de las tecnologías lingüísticas para satisfacer las necesidades de su población lingüísticamente compleja. Con casi el 30% de los ciudadanos de España que tienen como lengua materna una lengua regional distinta del castellano, el enfoque de la investigación debe tener en cuenta esta diversidad. El catalán, el vasco, el gallego y el gascón, en su variante aranesa, comparten oficialidad con el español en las regiones donde se hablan, mientras que tanto el aragonés como el asturiano están protegidos dentro de sus comunidades autónomas, aunque no son reconocidos como lenguas oficiales.

La descentralización política ha permitido a las distintas regiones españolas fomentar el desarrollo de la tecnología lingüística, con particular éxito en el caso del catalán y del vasco. El enfoque lingüístico intrínsecamente amplio de la investigación española, lejos de diluir el foco en las tecnologías para el español puede proporcionar un buen modelo para las políticas multilingües de Europa.

El compromiso de España con la investigación multilingüe es un modelo para la Unión Europea.

Por un lado existe una importante experiencia científica en idiomas minoritarios, como el catalán y el vasco, que es prácticamente única en Europa. La atención al desarrollo de productos en varios idiomas y servicios entre las lenguas regionales es poco común en otros estados miembros de la Unión. Por todos estos motivos, puede decirse que España es uno de los estados miembros con mayor potencial en tecnologías lingüísticas. Su compromiso con la investigación multilingüe, especialmente a través de programas regionales de I + DT, es un modelo para la Unión.

Por otro lado, la comunidad de investigadores españoles compite a nivel mundial en el desarrollo de productos para un idioma internacional como es el español. El español, como otras lenguas grandes, cuenta con herramientas de primera generación, por ejemplo de traducción automática, que se han desarrollado en EE.UU. o en otras partes de Europa, para aprovechar la oportunidad de mercado que representa una población de habla hispana de cuatrocientos o quinientos millones de personas.

## El cultivo del idioma en España

La Real Academia Española ejerce una influencia estandarizadora a través de la publicación de diccionarios y guías de gramática y estilo muy respetadas.

La Real Academia Española (RAE), fundada en 1713, junto con las academias de los 21 países de habla española, ejerce una influencia estandarizadora a través de la publicación de diccionarios y guías de gramática y estilo muy respetadas. Por cuestiones históricas de influencia y prestigio, existe una forma estandarizada de la lengua, principalmente basada en el español castellano, que es ampliamente reconocida para su uso en la literatura, los contextos académicos y los medios de comunicación.

La RAE, a través de su Fundación, ha creado dos grandes corpus textuales: el *Corpus de referencia del español actual* (CREA) y el *Corpus diacrónico del español* (CORDE). Estos corpus han sido utilizados por la Academia como una fuente de información para redactar diccionarios y gramáticas, tales como el *Diccionario del estudiante*, el *Diccionario panhispánico de dudas*, el *Diccionario esencial* y la *Nueva gramática de la lengua española*, que incorporan muchos ejemplos de la utilización real de la lengua tomados principalmente del CREA.

El Instituto Cervantes es una organización internacional sin ánimo de lucro creada por el gobierno español en 1991. Esta organización se ha expandido en más de 20 países con 54 centros dedicados a la cultura española e hispanoamericana y la lengua española. Los objetivos primordiales del Instituto son: promover la educación, el estudio y el uso universal del español como segunda lengua, dar apoyo a los métodos y actividades que contribuyan a la enseñanza del idioma español, y contribuir a la expansión de las culturas española e hispano-americana por los países de habla no española.

Desde 1997 hasta hoy, el Instituto Cervantes, junto con la Real Academia Española y la Asociación de Academias de la Lengua Española, han organizado cada tres años un Congreso Internacional de la Lengua Española. Este congreso es un foro internacional para discutir la situación, los problemas y los desafíos de la lengua española. Su objetivo es crear conciencia en los individuos, gobiernos e instituciones para la promoción de la lengua, su unidad y fomentar el diálogo entre la comunidad cultural hispánica.

La Fundéu BBVA (antes la Fundación del Español Urgente) [[8]](#endnote-8) es una organización sin ánimo de lucro creada en febrero de 2005 en Madrid. La Fundación fue creada en colaboración con la Real Academia Española, a partir del Departamento del Español Urgente de la Agencia EFE de noticias. En 2008 tomó el nombre de Fundéu BBVA. El objetivo principal de la Fundación es promover el buen uso del español en todos los medios de comunicación, incluyendo pero no exclusivamente, prensa, radio, televisión e Internet.

El objetivo principal de la Fundéu BBVA es promover el buen uso del español en todos los medios de comunicación, incluyendo prensa, radio, televisión e Internet.

La Fundación ha creado y mantiene una página web en la que se pueden encontrar los materiales elaborados por ella. El Manual del Español Urgente es una guía de referencia para aquellos que quieren escribir en español correcto, libre de modismos procedentes de otras lenguas. Un ejemplo del manual dice lo siguiente:

*Jugar un papel - la frase "Jugar un papel" aparece muy frecuentemente en las noticias y se utiliza cada vez más tanto en español escrito como hablado, ... "Jugar un papel" es un anglo-galicismo que tiene sus raíces en la expresión inglesa "to play a role" y en la francesa "jouer un rôle" ... en el sentido de "cumplir una misión o tarea” ... es una expresión inaceptable, ajena a nuestro idioma, como lo sería "jugar una comedia" o "jugar al piano".*

El manual también aborda otros aspectos lingüísticos, tales como la pronunciación, la ortografía o la puntuacion. Por ejemplo, con respecto a "seseo" y "ceceo" (las distintas formas locales de pronunciar 's', 'c' y 'z'), el Manual de Español Urgente dice:

*El seseo, es decir, la pronunciación de las letras c y z como s, es característico de algunas zonas del sur de la Península, Canarias y América. Esta pronunciación es apoyada por la norma culta. No así el ceceo (pronunciación de la s como z).*

En 2008, la Fundación del Español Urgente puso en marcha Wikilengua[[9]](#endnote-9), una wiki que sirve para consultar los usos aceptados del español. Construida a través de las aportaciones de la comunidad hispano parlante, se pueden consultar y corregir dudas frecuentes. Al ser abierta y accesible a todo el mundo, la Wikilengua sirve también de reflejo de la diversidad y riqueza del español en sus múltiples variedades. Este portal se desarrolla en torno a una comunidad de autores, traductores, editores, lingüistas, profesores, estudiantes, periodistas, y en general de todas las personas que están interesadas en el español, de forma individual o como parte de alguna entidad, y que desean compartir sus conocimientos con cientos de millones de hablantes de la lengua.

## 

## La lengua en la educación

El artículo 3 de la Constitución Española afirma que "el español es la lengua oficial del Estado. Todos los españoles tienen el deber de comprenderla y el derecho a usarla. Otras lenguas españolas también pueden ser lenguas oficiales dentro de sus comunidades autónomas, en función de sus respectivos estatutos. La rica diversidad de las lenguas de España constituye un patrimonio cultural que merece nuestro especial respeto y protección.” Cada comunidad “histórica” (Cataluña, País Vasco y Galicia), cuenta con leyes sobre el uso del español y de su idioma propio. Otras comunidades bilingües (Islas Baleares, Comunidad Valenciana, Navarra, Aragón y Asturias) han ido implementado diversas disposiciones legales para promover el uso de su lengua propia en el sistema educativo y la sociedad civil.

El español es la lengua oficial del Estado. Todos los españoles tienen el deber de comprenderla y el derecho a usarla.

Cada comunidad bilingüe cuenta con una Consejería de Política Lingüística para fomentar el uso de su lengua propia. Esta consejería propone, ejecuta y controla la aplicación de diversas medidas legales que se toman para promover las lenguas, y trabaja en estrecha colaboración con la Consejería de Educación de la comunidad. En el marco de la Ley de Educación del 3 de mayo de 2006 (LOE), se realizaron disposiciones legales para garantizar la enseñanza de la segunda lengua en las comunidades bilingües.

Cada comunidad bilingüe cuenta con una Consejería de Política Lingüística para fomentar el uso de su lengua propia.

En lo que se refiere a las disposiciones legales sobre los extranjeros, se estipula que todos los extranjeros menores de 18 años tienen el derecho (y el deber) de ser educados en las mismas condiciones que los españoles. Esto significa el acceso a una educación básica de 6 a 16 años de edad, que incluye la enseñanza primaria (6-12 años) y la educación secundaria obligatoria (12-16 años). Diferentes disposiciones favorecen que los jóvenes (dentro o fuera del sistema escolar) y los adultos aprendan las lenguas oficiales de España.

Diferentes disposiciones favorecen que los jóvenes y los adultos aprendan las lenguas oficiales de España.

Los programas relevantes de integración lingüística son responsabilidad de las comunidades autónomas y varían de una comunidad a otra. Sin embargo, estos programas se ofrecen de forma sistemática.

En el contexto de la descentralización del sistema educativo, la Ley de Educación (LOE), reafirma la responsabilidad compartida entre el Ministerio español de Educación y las Consejerías de Educación de las Comunidades Autónomas. En las comunidades monolingües, el Estado es responsable del 65% de la enseñanza y la Comunidad del 35%. En las comunidades bilingües, la responsabilidad del Estado se reduce al 55% y el de la Comunidad aumenta hasta el 45%. La enseñanza de una segunda lengua extranjera también está sujeta a las disposiciones adoptadas por las comunidades autónomas, sólo la primera lengua extranjera es obligatoria desde la escuela primaria y su estudio sigue siendo obligatorio hasta el final de la enseñanza secundaria (16 años) [[10]](#endnote-10) .

**Aspectos internacionales**

El español es el segundo idioma más estudiado en el mundo, después del inglés[[11]](#endnote-11), con cerca de dos millones de estudiantes de educación secundaria en la Unión Europea. También se ha convertido en el idioma extranjero más popular estudiado en los Estados Unidos (elegido por el 70% de los alumnos). En Brasil su estudio en la escuela secundaria es obligatorio. Además, el español es una lengua de cultura, dotado de una rica literatura.

El español es el segundo idioma más estudiado en el mundo, después del inglés, con cerca de dos millones de estudiantes de educación secundaria en la Unión Europea.

Sin embargo, a pesar del peso demográfico del español, de su posición como lengua de comunicación internacional y de la demanda actual del español como segunda lengua, su competitividad como lengua científica es seriamente cuestionada por el inglés. En el ámbito tecnológico, el inglés es claramente la lengua dominante. Su capacidad para generar nuevos términos técnicos y consolidar su uso es indiscutible. En el campo de las ciencias naturales y ambientales, el español es también prácticamente inexistente. Para hacer frente a esta situación, se han propuesto una serie de posibles soluciones. Estas incluyen el fortalecimiento de la enseñanza de ciencia y tecnología en español, la formación de estudiantes extranjeros en las universidades españolas, la organización de talleres regulares en escritura técnica y la creación de una agencia de noticias de ciencia en español.

En el ámbito tecnológico, el inglés es claramente la lengua dominante.

Por otro lado, el español como lengua de negocios está en condiciones de desafiar el dominio del inglés, por ser clave de un enorme mercado Latinoamericano. El valor económico de la lengua española ha sido incluso cuantificado[[12]](#endnote-12). Se considera que contribuye con más del 15% al PIB de España (por lo tanto, en 2009, más de 37.000 millones de euros) [[13]](#endnote-13).. Según una investigación realizada por la Fundación Telefónica, al usar el mismo idioma, se multiplica por dos o tres la cuota de mercado de las exportaciones de España a los países de habla española.

Se considera que el valor económico de la lengua española contribuye con más del 15% al PIB de España.

El español no es sólo un idioma en auge para los negocios, sino que es un negocio en sí mismo. Existe un nuevo tipo de turismo, formado por extranjeros que vienen a España para estudiar español, que ha aumentado desde 1995 a un ritmo de alrededor del 10% anual. En 2006, unos 150.000 turistas viajaron a España con este objetivo, y dejaron 255 millones de euros. El Instituto Cervantes tiene cursos específicos de español para los negocios, que parecen tener mucho éxito en los países de Europa del Este, debido a su relación con España, especialmente en la ámbito de la construcción y las obras públicas, y también debido a los mercados emergentes de Latinoamérica.

El Presidente del Consejo de Administración de INSEAD Business School (Universidad Autónoma de Madrid) prevé que en 2030 el español se convertirá en la segunda lengua de intercambio económico en el mundo, sólo por detrás del chino. Su uso generalizado requiere que se ofrezcan más másteres de negocios y cursos de capacitación en este idioma. Los métodos de enseñanza tradicionales están empezando a dar paso a las nuevas tecnologías y al e-learning. Hasta el momento el mayor número de estudiantes interesados ​​ha venido de países de América Latina. Sin embargo, se están incorporando cada vez más estudiantes de otros orígenes, especialmente los profesionales de habla portuguesa, además de alemanes y franceses. Los estudiantes que buscan este tipo de cursos son profesionales con experiencia que desean avanzar en sus carreras, así como las empresas que promueven la formación de su personal con el fin de lograr una mayor competitividad.

El español es reconocido como uno de los idiomas oficiales de las Naciones Unidas, la Unión Europea, la Organización de Estados Americanos, la Organización de Estados Iberoamericanos, la Unión Africana, la Unión de Naciones Suramericanas, la Unión Latina, y la Caricom, y tiene un estatus legal en el Acuerdo de Libre Comercio de Norteamérica.

El español es lengua oficial de la UE pero no es un idioma de trabajo. El Gobierno español también ha hecho esfuerzos a favor de incluir el catalán, el gallego y el vasco, entre las lenguas oficiales de las instituciones europeas. Actualmente estas se consideran semi-oficiales, junto con el escocés, el gaélico y el galés. Al no ser idioma de trabajo, hay una gran cantidad de documentos que no están traducidos al español y una cantidad aún mayor no traducida al resto de las lenguas de España. La mayoría de estos documentos están disponibles solamente en inglés, algunos de ellos están también en francés o en alemán. Lo mismo ocurre con la mayoría de las páginas web de la UE.

La tecnología lingüística puede ayudar a disminuir los inconvenientes personales y económicos a los que se enfrentan de forma cotidiana los hablantes no nativos de inglés.

La tecnología lingüística puede ayudar a abordar este reto ofreciendo servicios tales como la traducción automática o la recuperación de información multilingüe y por lo tanto ayudar a disminuir los inconvenientes personales y económicos a los que se enfrentan de forma cotidiana los hablantes no nativos de inglés.

## El español en Internet[[14]](#endnote-14)

Con frecuencia se afirma que el inglés domina el mundo de la informática y de Internet, y que aquellos que deseen utilizarla deben primero aprender inglés. Esto puede haber sido cierto en los primeros días de la tecnología, pero el desconocimiento del inglés ya no es una barrera infranqueable. Lo que comenzó como un fenómeno de habla inglesa se ​​ha convertido rápidamente en un asunto multilingüe. Muchas páginas web corporativas ahora emplean estrategias multilingües que convierten la elección del idioma en una preferencia de usuario. La traducción automática de los contenidos web está sólo a un clic de distancia. Además, Internet está demostrando ser un recurso muy útil para aquellos interesados ​​en aprender lenguas minoritarias. Un análisis publicado en noviembre de 2005 por Byte Level Research concluyó que "la próxima revolución de Internet no será en inglés". Al tiempo que el inglés no deja de ser importante en Internet, otros idiomas, como el chino, el ruso, el español y el portugués, son cada vez más importantes[[15]](#endnote-15).

El número total de usuarios de Internet en los diferentes países de habla española fue de más de 153 millones en 2010, con un 20,8% procedentes de España.

La próxima revolución de Internet no será en inglés.

Las estadísticas globales de uso de Internet para 2007 muestran el español como al tercer idioma más utilizado en Internet, después del inglés y del chino[[16]](#endnote-16). El número total de usuarios de Internet en los diferentes países de habla española fue de más de 153 millones en 2010, con un 20,8% procedentes de España. La penetración de Internet en España alcanzó el 62,6% en el 2010 según Nielsen Online. [[17]](#endnote-17)

El estudio “Perfil Sociodemográfico de los Internautas – Análisis de los datos del INE de 2010”, elaborado por el Observatorio Nacional de las Telecomunicaciones y la Sociedad de la Información (ONTSI) [[18]](#endnote-18), se basa en los datos del Instituto Nacional de Estadística (INE) en su “Encuesta sobre equipamiento y uso de tecnologías de la información y comunicación en los hogares 2010” (TIC-H 2010), y señala que el uso de Internet es cada vez más intensivo:

* El 44,4% de la población española de 16 a 74 años, es decir, 15,4 millones de personas, utiliza Internet diariamente.
* Casi 27 millones de ciudadanos españoles de 10 y más años han utilizado Internet en alguna ocasión.
* La diferencia de uso de Internet entre hombres y mujeres se ha visto reducida en los últimos años, principalmente entre los usuarios que acceden a la Red todas las semanas.
* La edad es una variable determinante: al disminuir la edad, aumenta el porcentaje de internautas. Por otro lado, prácticamente la totalidad de los estudiantes ha accedido a la Red en alguna ocasión.
* El nivel de estudios terminados marca diferencias del uso realizado de la Red, cuanto mayor es el nivel de estudios, mayor es el porcentaje de internautas.

Otro informe elaborado por el ONTSI, basado en datos extraídos de una encuesta sobre el uso de TIC en las empresas, también llevado a cabo por el INE, afirma que el 100% de las grandes empresas disponen de ordenadores, acceso a Internet y correo electrónico. También señala que 2 de cada 3 microempresas disponían de teléfonos móviles y ordenadores en el primer trimestre de 2010. Más del 90% de las empresas grandes y medianas y casi el 40% de las microempresas interactuan con la administración pública a través de Internet. Las dos razones principales para la interacción son la obtención de información y la descarga de formularios. El crecimiento más notable en las PYME y las grandes empresas se ha observado en la gestión electrónica completa, que se ha incrementado en 5,2 puntos respecto al año anterior.

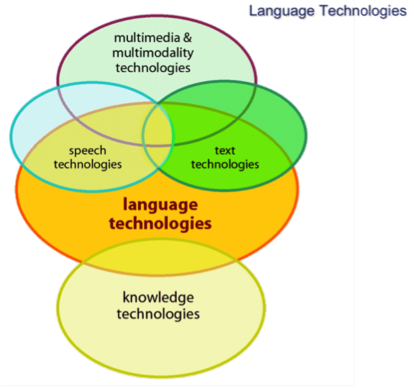
Todos los sectores tienen tasas de penetración de Internet superior al 90%, y el sector financiero del 100%. El 97,2% de las PYME y de las grandes empresas tienen acceso a Internet. La búsqueda de información y el uso de servicios financieros y bancarios son las actividades más comunes para las empresas con Internet, con porcentajes del 96,4% y el 90,2%, respectivamente. La Web se utiliza como una plataforma de comunicación en el 86,8% de las ocasiones. En 2010, el 58,1% de las microempresas tenía acceso a Internet, dos puntos más que el año anterior, acumulando un crecimiento de 5 puntos en dos años. El 75,5% de las microempresas con acceso a Internet lo utilizan como una plataforma de comunicación, y el 69,3% lo utiliza para acceder a servicios bancarios y financieros.

Todos los sectores tienen tasas de penetración de Internet superior al 90%. La búsqueda de información y el uso de servicios financieros y bancarios son las actividades más comunes para las empresas con Internet.

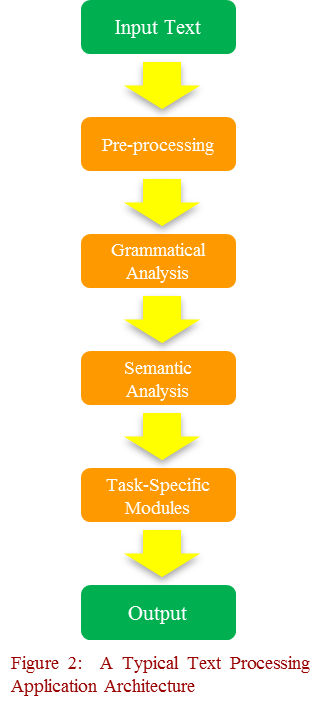
Para la tecnología lingüística, Internet es importante en dos sentidos. Por un lado, la gran cantidad de datos lingüísticos disponible en versión digital representa una rica fuente de información para el análisis del uso del lenguaje natural mediante técnicas estadísticas. Por otro lado, Internet ofrece una amplia gama de áreas de aplicación de la tecnología lingüística. [[19]](#endnote-19).

La aplicación web más utilizada son sin duda los motores de búsqueda, lo que implica el procesamiento automático del lenguaje en múltiples niveles, como veremos con más detalle en la siguiente sección. La búsqueda requiere tecnología lingüística sofisticada, diferente para cada idioma. Para una lengua flexiva como el español, un lematizador, capaz de generalizar a partir de las formas flexionadas, es una herramienta muy importante. Sin embargo, los usuarios de Internet y proveedores de contenido de la web también se pueden beneficiar de la tecnología lingüística de maneras menos obvias, por ejemplo, al traducir de forma automática el contenido web de un idioma a otro. Teniendo en cuenta los altos costos asociados con la traducción manual de estos contenidos, resulta sorprendente lo poco que se invierte en tecnología lingüística que pueda hacer frente a estas necesidades. En el siguiente capítulo, presentamos una introducción a las tecnologías lingüísticas y sus áreas de aplicación principales, así como una evaluación de la situación actual de los recursos y las herramientas de tecnología lingüísticas para el español.

# Recursos de tecnología lingüística para el español

Las tecnologías lingüísticas son programas diseñados para tratar con el lenguaje humano (o lenguaje natural, por oposición a lenguaje formal) y por lo tanto, a menudo se las conoce como "tecnologías del lenguaje natural". El lenguaje humano se presenta tanto en forma oral como escrita. Mientras que el habla es la forma de comunicación lingüística más antigua y, en términos evolutivos, la más natural, la información más compleja y la mayor parte del conocimiento humano se almacena y se transmite a través de textos escritos. Las tecnologías lingüísticas relacionan la lengua con diversas formas de conocimiento, con independencia del medio en que esta se expresa (voz o texto). La figura de la derecha ilustra el panorama general de las tecnologías lingüísticas. Cuando nos comunicamos, el lenguaje se combina con otros modos de comunicación y medios de información; por ejemplo, hablar puede implicar gestos y expresiones faciales. Los textos digitales se enlazan con imágenes y sonidos. Las películas pueden contener lenguaje en forma hablada y escrita. En otras palabras, las tecnologías del habla y de texto se superponen e interactúan con otras tecnologías que facilitan el procesamiento de la comunicación multimodal y de los documentos multimedia.

En las siguientes secciones, discutiremos los principales ámbitos de aplicación de la tecnología lingüística: la corrección automática, la búsqueda en la Web, las tecnologías del habla y la traducción automática. Esto incluye aplicaciones y tecnologías básicas, como

* corrección ortográfica
* asistencia a la edición
* aprendizaje de idiomas asistido por ordenador
* recuperación de información
* extracción de información
* resumen de texto
* búsqueda de respuestas
* reconocimiento de voz
* síntesis de voz

Antes de entrar en detalle en cada una de estas áreas, describiremos brevemente la arquitectura de un sistema de procesamiento típico.

## Arquitectura de las aplicaciones

Los diferentes módulos que constituyen las aplicaciones de procesamiento del lenguaje suelen reflejar los diferentes niveles de análisis lingüístico. La figura de la derecha muestra una arquitectura muy simplificada de un sistema de procesamiento de textos típico. Los tres primeros módulos gestionan la estructura y el significado del texto de entrada:

1. Pre-procesamiento: limpia los datos, analiza o suprime el formato de entrada, detecta el idioma del texto, lo normaliza, etc.
2. Análisis gramatical: identificación del predicado y de sus argumentos, modificadores y otras partes del discurso, así como de la estructura de la oración.
3. Análisis semántico: desambigua (es decir, calcula el significado apropiado de las palabras en un contexto determinado), resuelve las anáforas (es decir, identifica los nombres a los que se refieren los pronombres en la oración), y representa el significado de la frase en una forma interpretable por la máquina.

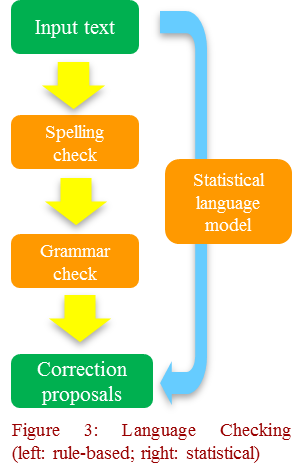
Después de analizar el texto, los módulos específicos pueden realizar otras operaciones, tales como resumen automático o consulta de bases de datos. Esta es una descripción simplificada e idealizada de la arquitectura típica de las aplicaciones lingüísticas y nos sirve para ilustrar la complejidad de estas.

Después de una breve introducción a cada una de las principales áreas de aplicación de las tecnologías lingüísticas, ofrecemos un breve resumen del estado de la investigación y de la educación en tecnología lingüística en España, y terminamos con un repaso de los programas de investigación, pasados y presentes. A continuación, presentamos una estimación experta de las herramientas y los recursos básicos del español según una diversidad de criterios, como por ejemplo, su disponibilidad, madurez y calidad. La situación general de las tecnologías lingüísticas en lengua española se resume en una tabla.

## Áreas principales de aplicación

En esta sección nos centramos en las herramientas y los recursos lingüísticos más importantes, y damos una visión general de las actividades de TL en España.

### Corrección de textos

 Cualquiera que haya usado un procesador de texto como Microsoft Word sabe que tiene un corrector ortográfico que resalta los errores de ortografía y propone correcciones. Los primeros programas de corrección ortográfica comparaban cada palabra con un diccionario de palabras escritas correctamente. En la actualidad estos programas son mucho más sofisticados. Mediante algoritmos de análisis de texto, específicos para cada lengua, son capaces de detectar errores morfológicos (por ejemplo, la formación del plural), así como errores de sintaxis, tales como la falta de concordancia entre sujeto y verbo (por ejemplo, \**ella escriben una carta*).

El tratamiento de muchos tipos de errores requiere un análisis del contexto. Un caso típico es cuando el error ortográfico transforma una palabra en otra palabra existente. En el siguiente ejemplo, la primera frase contiene un par de errores comunes (problemas con los acentos ortográficos y la omisión de /h/ muda). La segunda frase es la versión corregida de la primera:

Mí calculo es que hoy a venido mas publico que ayer.

Mi cálculo es que hoy ha venido más público que ayer.

Para analizar el contexto lingüístico son necesarias reglas gramaticales laboriosamente codificadas por expertos, o bien un modelo de lenguaje estadístico construido a partir de corpus. En este último caso, el modelo calcula la probabilidad de que una palabra concreta aparezca en una posición específica (por ejemplo, entre las palabras que la preceden y la siguen). Por ejemplo, [Mi cálculo] o [ha venido] son secuencias de palabras mucho más probables que [Mí calculo] o [a venido] respectivamente. Un modelo estadístico de lenguaje se construye automáticamente a partir de una gran cantidad de texto (correcto), lo que se conoce como corpus textual. La mayor parte de las aplicaciones lingüísticas, ya sean basadas en reglas o en modelos estadísticos, se han desarrollado para el inglés. Otras lenguas han podido beneficiarse de toda la investigación realizada en torno al inglés. Sin embargo, no todos los tratamientos desarrollados para esta lengua se aplican bien a otras. Por ejemplo, los idiomas con una morfología más rica, o con un orden de palabras flexible, como el español, pueden necesitar un análisis lingüístico más profundo para alcanzar el mismo grado de precisión en la corrección de los textos.

La corrección de textos también se utiliza en sistemas de creación controlada de textos, es decir, entornos para la edición de manuales y documentación que siguen normas de escritura especiales, en el ámbito de la salud, ingeniería y otros. Para evitar reclamaciones sobre el uso incorrecto de sus aparatos o por daños resultantes de instrucciones mal entendidas, las empresas están insistiendo cada vez más en la calidad de la documentación técnica, y de las traducciones de sus versiones multilingües. Los avances en el procesamiento del lenguaje natural han llevado al desarrollo de software de edición que ayuda al autor de documentación técnica a usar vocabulario especializado y estructura de las frases compatibles con las normas de la industria y que se ajustan a las restricciones terminológicas corporativas.

Sólo una pocas empresas españolas ofrecen productos en este área. Por ejemplo, Daedalus, que nació como una spin-off de dos grupos de investigación de la Universidad Politécnica de Madrid (UPM) y de la Universidad Autónoma de Madrid (UAM), ha desarrollado STYLUS, un corrector ortográfico, gramatical y de estilo en español. La tecnología está disponible gratuitamente a través de su portal web, pero también puede integrarse en cualquier herramienta de gestión de contenidos. WinCorrect es un corrector ortográfico y gramatical desarrollado por Maxigramar (también con una versión en línea gratuita) y Signum es una PYME basada en Ecuador que desarrolló el corrector ortográfico para el español licenciado por Microsoft para Office.

Además de los correctores ortográficos y los servicios de edición, la tecnología de corrección de textos es también importante en el campo del aprendizaje de lenguas asistido por ordenador. Los motores de búsqueda, como Google, también utilizan correctores que corrigen de forma automática las consultas y proponen alternativas.

### Búsquedas en la Web

Buscar en la Web, en las intranets o en las bibliotecas digitales es probablemente la aplicación de TL más utilizada en la actualidad. El motor de búsqueda de Google, que comenzó en 1998, ahora se encarga de aproximadamente del 80% de todas las consultas. La interfaz de búsqueda de Google y la página donde muestra los resultados no han cambiado significativamente desde la primera versión. Sin embargo, en la versión actual, Google ofrece corrección de ortografía de las palabras mal escritas, y recientemente ha incorporado capacidades básicas de búsqueda semántica que pueden mejorar la precisión de la búsqueda mediante el análisis del significado de los términos de la consulta. El éxito de Google muestra que una gran cantidad de datos y unas técnicas eficientes de indexación pueden dar resultados satisfactorios, basados únicamente en métodos estadísticos.

Para requerimientos de información más sofisticados, es esencial integrar conocimiento lingüístico más complejo. El uso de recursos léxicos como tesauros (por ejemplo, WordNet) u ontologías aumenta la capacidad de búsqueda de páginas usando sinónimos de los términos originales, como por ejemplo "Energía Atómica" y "Energía nuclear", o incluso términos más vagamente relacionados.

La próxima generación de los motores de búsqueda tendrá que incluir tecnología lingüística mucho más sofisticada, con el fin de dar respuesta a las consultas que consisten en una pregunta o una petición, en lugar de una lista de palabras clave. Para la consulta, "Dame una lista de todas las empresas que fueron adquiridas por otras compañías en los últimos cinco años," el sistema TL tiene que analizar la frase sintáctica y semánticamente, así como proporcionar un índice para recuperar rápidamente los documentos pertinentes. Una respuesta satisfactoria requieren un análisis sintáctico para analizar la estructura gramatical de la oración y determinar que el usuario quiere las compañías que han sido adquiridos, no las empresas que adquirieron otras empresas. Para la expresión “últimos cinco años”, el sistema necesita determinar los años pertinentes. Y, la consulta tiene que ser comparada con una enorme cantidad de datos no estructurados con objeto de encontrar la información relevante que el usuario desea. Esto se llama "recuperación de información", y consiste en la búsqueda y clasificación de los documentos pertinentes. Para generar una lista de empresas, el sistema también tiene que reconocer una secuencia particular de palabras en un documento como nombres de empresa, un proceso llamado "reconocimiento de entidades con nombre".

La recuperación de documentos en varios idiomas a partir de una única consulta en otro idioma distinto añade dificultad a la tarea. La recuperación de información multilingüe implica traducir automáticamente la consulta a todos los idiomas destino que sea posible y luego traducir los resultados de nuevo a la lengua origen.

Ahora que cada vez es más frecuente encontrar datos en formatos no textuales, se precisan servicios de recuperación de información multimedia que busquen imágenes, archivos de audio y de video. En el caso de archivos de audio y video, un módulo de reconocimiento de voz debe convertir el contenido de la voz en texto (o en una representación fonética) que pueda ser comparada con la consulta del usuario.

En España, unas pocas empresas desarrollan tecnología lingüística dirigida a la búsqueda y recuperación de información multilingüe, tanto desde Internet como desde sistemas de información internos. Entre las principales, se encuentran las siguientes: iSOCO, Dédalo, Inbenta, Bitext y Thera, esta última también spin-off, en este caso de la Universitat de Barcelona. Su tecnología incorpora herramientas de traducción automática, así como componentes de reconocimiento de entidades nombradas, búsqueda difusa y etiquetado semántico.

Estas empresas centran su desarrollo en el suministrar motores de búsqueda avanzados para portales de interés especial mediante el uso de semántica específica a un dominio. Debido a la gran demanda de potencia de procesamiento, estos motores de búsqueda sólo son eficientes en la gestión de corpus relativamente pequeños. El tiempo de procesamiento es varias miles de veces mayor que la que necesita un motor de búsqueda estándar estadístico como Google. Estos motores de búsqueda tienen una alta demanda para el modelado de dominio específico, pero no se pueden utilizar en la Web con sus miles y miles de millones de documentos.

### Procesamiento del habla

La tecnología del procesamiento de voz se utiliza para crear interfaces de lengua hablada.

La tecnología del procesamiento de voz se utiliza para crear interfaces que permiten a los usuarios interactuar a través de lengua hablada en lugar de una pantalla gráfica, teclado y ratón. Hoy en día, las interfaces de usuario de voz se utilizan generalmente en los servicios de telefonía (semi-)automatizada proporcionada por las empresas a los clientes, empleados o socios. Los ámbitos comerciales que dependen en gran medida de este tipo de interfaces incluyen la banca, los suministros de servicios, el transporte público y las telecomunicaciones. Otros usos de la tecnología de voz incluyen interfaces para sistemas de navegación y el uso de la lengua oral como alternativa a las interfaces gráficas o la pantalla táctil de los smartphones.

Las tecnologías del habla comprenden las siguientes:

1. El reconocimiento automático de voz, identifica las palabras a partir de una determinada secuencia de sonidos emitidos por un usuario.
2. La comprensión del lenguaje natural, analiza la estructura sintáctica de la expresión de un usuario y lo interpreta de acuerdo con el sistema en cuestión.
3. La gestión del diálogo, determina qué acción tomar, dada la entrada del usuario y la funcionalidad del sistema.
4. La síntesis de voz (texto a voz) transforma la respuesta del sistema en sonidos para el usuario.

Uno de los principales retos de los sistemas de reconocimiento es reconocer con precisión las palabras que un usuario emite. Esto implica restringir la gama de expresiones posibles de los usuarios a un conjunto limitado de palabras clave, o bien construir manualmente modelos de lenguaje que cubren una amplia gama de expresiones del lenguaje natural. Utilizando técnicas de aprendizaje automático, también se pueden generar modelos de lenguaje automáticamente a partir de corpus del habla, es decir, de grandes colecciones de archivos de audio, acompañadas de transcripciones. Restringir la gama de expresiones por lo general obliga a los usuarios a utilizar una interfaz de voz de una manera rígida y puede no ser bien aceptado por el usuario, pero la creación y mantenimiento de modelos ricos de lengua aumenta significativamente los costos del sistema.

Las empresas tienden a utilizar frases pregrabadas por locutores profesionales para generar la salida de la interfaz de usuario de voz. Para expresiones estáticas donde el texto no depende de los contextos particulares de uso o de los datos personales del usuario, esto puede ser suficiente. Pero en caso de contenidos más dinámicos, este sistema puede generar una entonación poco natural. Los sistemas actuales de texto a voz trabajan sobre este problema y son cada vez mejores en la producción de un sonido natural de expresiones generadas de forma dinámica por el sistema.

Las interfaces de voz que se hallan en el mercado, se han estandarizado considerablemente durante la última década, en términos de sus diversos componentes tecnológicos. También se ha consolidado el mercado en el reconocimiento y en la síntesis de voz. Los mercados nacionales de los países del G20 (países económicamente resistente con alta población) ha estado dominado por sólo cinco actores globales, con Nuance (EE.UU.) y Loquendo (Italia) entre los más importantes de Europa, también para el español, aunque algunas pequeñas compañías locales están empezando a competir, como Verbio, que es un spin-off de la Universitat Politècnica de Catalunya y dispone de tecnología de voz propia.

En cuanto a la tecnología de gestión de diálogo, los mercados están fuertemente dominados por actores nacionales, que suelen ser pequeñas empresas.

La mayoría de las empresas en el mercado español de las tecnologías de la voz son esencialmente desarrolladores de aplicaciones. Los principales actores en el mercado español son los siguientes: Indsys (Sistemas Inteligentes de diálogo), Fonetic , Ydilo y NaturalVoz. En lugar de depender de un modelo de negocio basado en las licencias de software, estas empresas se han posicionado principalmente como proveedores de servicios que crean interfaces de usuario de voz como parte de un servicio de integración de sistemas. En el área de la interacción del habla, no existe aún un mercado real para las tecnologías de análisis basadas en el análisi sintáctico y semántico.

La demanda de interfaces de voz en España ha crecido rápidamente en los últimos cinco años, impulsado por la creciente demanda de atención automática a los clientes, optimización de costes de los servicios telefónicos automatizados, y la creciente aceptación de la lengua oral como medio para la interacción hombre-máquina.

Mirando hacia el futuro, la propagación de los teléfonos inteligentes, o smartphones, como nueva plataforma para la gestión de relaciones con los clientes, además de la telefonía fija, Internet y correo electrónico, supondrá cambios significativos. Esto también afectará a como se usa la tecnología de voz. A la larga, el lenguaje hablado tendrá un papel mucho más importante como una entrada fácil de usar para los teléfonos inteligentes. Esto se deberá principalmente a mejoras en el reconocimiento de voz independiente de locutor, a través de servicios de dictado que ya se ofrecen como servicios centralizados a los usuarios de los smartphones.

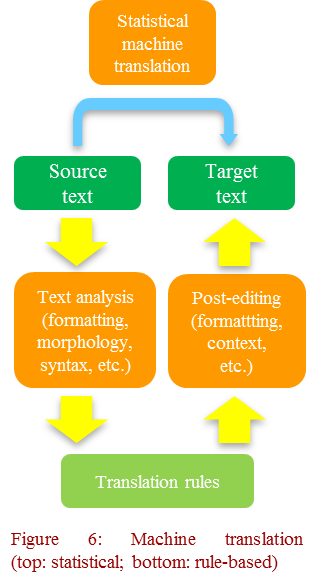
### Traducción Automática

La aproximación más simple a la traducción automática consiste en reemplazar automáticamente las palabras de un texto en un idioma, por palabras en otro idioma.

La idea de utilizar ordenadores para traducir de una lengua a otra se remonta a 1946 y la investigación en este ámbito gozó de una importante financiación durante la década de 1950 y de nuevo en la década de 1980. Sin embargo, la traducción automática (TA) aún no es capaz de cumplir su objetivo inicial de traducción automática de gran calidad e independiente de dominio.

La aproximación más simple a la traducción automática consiste en reemplazar automáticamente las palabras de un texto en un idioma, por palabras en otro idioma. Esto puede ser útil en ámbitos que tienen un lenguaje muy restringido, como los informes metorológicos. Sin embargo, para producir una buena traducción de textos menos estandarizados, hay que asociar unidades mayores de texto (sintagmas, oraciones, o incluso párrafos completos) con sus traducciones en la lengua destino. La mayor dificultad es que el lenguaje humano es ambiguo. La ambigüedad afecta múltiples niveles, por ejemplo la desambiguación de sentidos a nivel léxico (un jaguar es una marca de coche o un animal) o la asignación de casos a nivel sintáctico, por ejemplo:

The woman saw the car and her husband, too.

[La mujer vio el coche y su marido también.]

[La mujer vio el coche y **a** su marido también..]

Una forma de construir un sistema de TA es escribir reglas lingüísticas. Los sistemas basados ​​en reglas (o basados en conocimiento lingüístico) analizan el texto de entrada y crean una representación simbólica intermedia a partir de la cual puede generarse el texto en la lengua de destino. El éxito de estos métodos depende en gran medida de la disponibilidad de léxicos amplios con información morfológica, sintáctica y semántica, y de grandes conjuntos de reglas gramaticales cuidadosamente diseñadas por expertos lingüistas. Este es un proceso muy largo y costoso.

A finales de 1980, cuando aumentó la potencia de cálculo de los ordenadores y se hizo más barata, creció el interés por la traducción automática basada en modelos estadísticos. Los modelos estadísticos se obtienen a partir del análisis de textos bilingües, como el corpus Europarl, que contiene las actas del Parlamento Europeo en 11 idiomas distintos. Si se cuenta con una cantidad suficiente de datos, la TA estadística (o basada en los datos) funciona lo suficientemente bien como para obtener una traducción aproximada de un texto mediante el procesamiento de versiones paralelas y la búsqueda de emparejamientos potenciales de palabras. Pero, a diferencia de los sistemas basados en conocimiento, la TA estadística a menudo genera resultados agramaticales. La TA estadística es ventajosa porque requiere menos esfuerzo humano, y también puede cubrir algunas particularidades de la lengua (por ejemplo, expresiones, idiomáticas) que suelen ser ignoradas por los sistemas basados en el conocimiento.

Las virtudes y los defectos de la TA basada en reglas y de la TA estadística tienden a ser complementarios, de modo que en la actualidad los investigadores se centran en enfoques híbridos que combinan ambas metodologías. Una posibilidad consiste en utilizar los dos tipos de sistemas junto con un módulo de selección que decide sobre el mejor resultado para cada oración. Sin embargo, los resultados para las frases de más de 12 palabras no suelen ser buenos. Una solución mejor consiste en combinar las mejores traducciones de cada parte de la frase a partir de múltiples resultados, lo que puede resultar bastante complejo, ya que la segmentación y alineamiento de las distintas partes no suele ser obvia.

Lucy Software es una empresa internacional, líder en TA, que cuenta con una importante filial en España, Lucy Ibérica, anteriormente Translendium. Lucy Ibérica es responsable del desarrollo de los pares de idiomas que incluyen el español y todos los pares de idiomas que incluyen cualquier otra lengua ibérica (catalán, portugués, gallego y vasco). Word Magic es un conocido sistema de TA Español-Inglés (y viceversa) desarrollado por una empresa estadounidense. Tanto Lucy como Word Magic son sistemas basados en reglas gramticales. Si bien se investiga en sistemas de TA estadísticos e híbridos tanto en el contexto nacional como internacional, esta metodología ha tenido menos éxito en los negocios que en la investigación hasta el momento, con sólo unas pocas empresas que ofrece TA estadística adaptada al cliente, como Pangeanic.

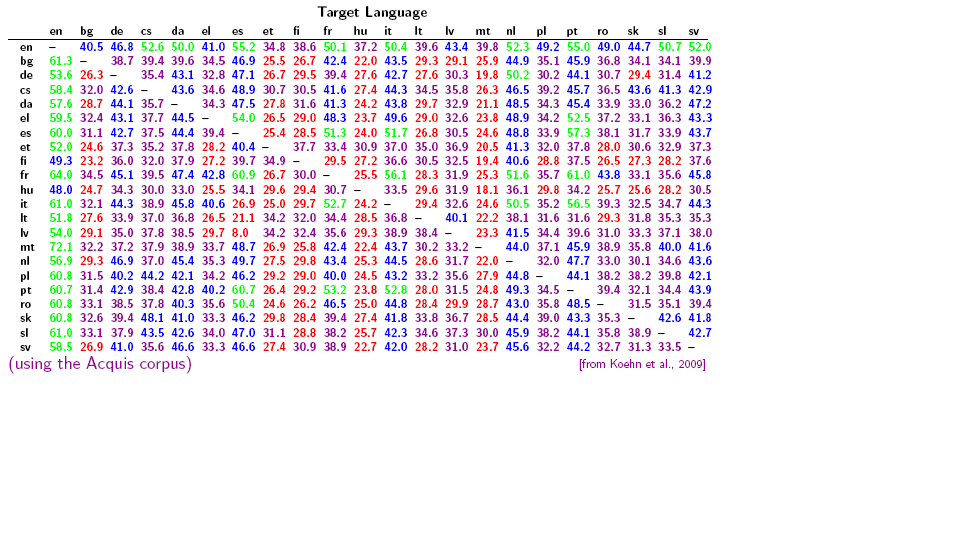
Apertium es una plataforma de traducción automática, libre y de código abierto, que proporciona un motor de traducción, independiente de lengua, diseñado inicialmente por el grupo Transducens de la Universitat d'Alacant y, posteriormente, desarrollado en el marco del proyecto Opentrad , financiado con fondos nacionales. Entre los sistemas de traducción actuales que usan la tecnología Apertium, encontramos interNOSTRUM (español-catalán), Traductor Universia (español-portugués) y Matxin (euskera-español), los dos primeros desarrollados por Transducens y el tercero por el grupo IXA de la Euskal Herriko Unibertsitatea. Es posible usar Apertium para construir sistemas de traducción automática para una gran diversidad de combinaciones de idiomas (hay más de 20 hasta la fecha). Apertium usa formatos simples basados ​​en XML para codificar los datos lingüísticos necesarios (ya sea a mano o por conversión de datos preexistentes), que se compilan utilizando las herramientas proporcionadas por el sistema y se convierten a los formatos de alta velocidad utilizado por el motor de traducción.

El uso de la traducción automática puede aumentar significativamente la productividad siempre que el sistema se adapte de forma inteligente a la terminología específica del usuario y se integre en un flujo de trabajo.

Todavía hay un enorme potencial para mejorar la calidad de los sistemas de MT. Los desafíos implican la adaptación de los recursos lingüísticos a un dominio determinado o área de usuario, y la integración de la tecnología en los flujos de trabajo que ya cuentan con bases terminológicas y las memorias de traducción. Otro problema es que la mayoría de los sistemas actuales están centrados en el inglés y muy pocos proporcionan traducción entre el español y otras lenguas, distintas del inglés.

Existen campañas de evaluación periódicas que ayudan a comparar la calidad de los sistemas de MT existentes y la cobertura de los distintos pares de idiomas. La tabla de abajo, que fue establecida en el proyecto de la CE Euromatrix +, muestra los resultados de los distintos pares para 22 de las 23 lenguas oficiales de la UE (el irlandés Gaélico no se comparó). Los resultados se clasifican de acuerdo a la métrica BLEU, que indica una puntuación más alta para una mejor traducción. (Un traductor humano alcanzaría una puntuación de unos 80 puntos.)

Los mejores resultados (en verde y azul) corresponden a las lenguas que se benefician de un mayor esfuerzo de investigación en programas coordinados y de la existencia de corpus paralelos (por ejemplo, inglés, francés, holandés, español y alemán). Los idiomas con los peores resultados (en rojo), o bien carecen de dichos esfuerzos de investigación, o bien se trata de idiomas tipológicamente muy distantes al resto (por ejemplo, húngaro, maltés y finés).



Rendimiento de los sistemas de Traducción Automática para los pares de lenguas del proyecto Euromatrix+

## Otras áreas de aplicación

Las tecnologías lingüísticas proporcionan funcionalidades importantes de forma no siempre evidente.

La construcción de aplicaciones lingüísticas implica un conjunto de tecnologías que no siempre aparecen de forma evidente en el nivel de interacción con el usuario, pero que proporcionan funcionalidades importantes al sistema en cuestión. Todos ellos constituyen temas importantes de investigación que han evolucionado hasta convertirse en sub-disciplinas específicas de la lingüística computacional.

La **búsqueda de respuestas** o *question answering*, por ejemplo, es un área activa de investigación para la que se han construido corpus anotados y se han puesto en marcha concursos científicos. El concepto de búsqueda de respuestas va más allá de las búsquedas convencionales basadas en palabras clave (en las que el motor de búsqueda responde mediante la entrega de una colección de documentos potencialmente relevantes). Este tipo de búsqueda permite a los usuarios hacer una pregunta concreta a la que el sistema proporciona una respuesta única. Por ejemplo:

Pregunta: ¿Qué edad tenía Neil Armstrong cuando pisó la luna?

Respuesta: 38.

A pesar de que la búsqueda de respuestas está obviamente relacionada con la búsqueda en la Web, hoy en día es un término general que engloba temas de investigación tales como: qué tipos de preguntas hay y cómo deberían tratarse; cómo pueden analizarse y compararse un conjunto de documentos que podrían contener la respuesta (¿ofrecen respuestas contradictorias?), y cómo puede extraerse de forma fiable la información específica (es decir, la respuesta) de un documento sin descuidar el contexto.

Esto a su vez está relacionado con la **extracción de información**, un ámbito que se hizo muy popular e influyente cuando la lingüística computacional dio un giro hacia la estadística en la década de 1990. La extracción de información tiene como objetivo identificar información específica en clases de documentos específicos, como por ejemplo, la detección de los actores clave en la adquisición de empresas a partir de informaciones periodísticas. Otra situación común que se ha estudiado son los informes sobre incidentes terroristas. El problema aquí consiste en proyectar el texto hacia una plantilla que especifique el autor, el objetivo, hora, lugar y los resultados del incidente. La extracción de información consiste básicamente en llenar plantillas relativas a dominios específicos, por lo que constituye otro ejemplo de tecnología “entre bastidores” que en la práctica tiene que ser integrada en un entorno de aplicación adecuada.

El **resumen** y la **generación de texto** son dos ámbitos muy relacionados que pueden actuar como aplicaciones independientes o desempeñar un papel de apoyo "entre bastidores". El resumen de textos intenta dar los elementos esenciales de un texto largo de una forma breve, y es una de las funcionalidades disponibles en Microsoft Word. Utiliza principalmente métodos estadísticos para identificar las palabras "importantes" de un texto (es decir, palabras que son muy frecuentes en el texto en cuestión, pero menos frecuentes en el uso del lenguaje en general) y determinar así qué frases contienen la mayor parte de estas palabras "importantes". Estas frases luego se extraen y combinan para crear el resumen. En este escenario comercial, muy común, el resumen es simplemente una forma de extracción de frases, y el texto se reduce a un subconjunto de las mismas. Un enfoque alternativo, que también se ha investigado, consiste en generar oraciones completamente nuevas que no existían en el texto original. Esto requiere una comprensión más profunda del texto, lo que significa que hasta el momento este enfoque es mucho menos robusto. En general, un generador de texto rara vez se utiliza como una aplicación independiente, sino que se integra en un entorno de software más grande, como, por ejemplo, un sistema de información clínica que recoge, almacena y procesa los datos de los pacientes. La creación de informes es sólo una de las muchas aplicaciones del resumen de texto.

La investigación en estas tecnologías está mucho menos desarro-llada para la lengua española que para el inglés.

La investigación en estas tecnologías está mucho menos desarrollada para la lengua española que para el inglés. La búsqueda de respuestas, la extracción de información y el resumen han sido objeto de numerosas competiciones en los EE.UU. desde la década de 1990, organizados principalmente por las organizaciones gubernamentales DARPA y NIST. Estas competiciones han mejorado significativamente el nivel de las tecnologías, pero se han centrado sobre todo en el inglés. Como resultado, casi no hay corpus anotado u otros recursos especiales necesarios para realizar estas tareas en español. Los sistemas de resumen que utilizan métodos puramente estadísticos son en gran parte independientes de idioma y existe una serie de prototipos de investigación disponibles. En la generación de texto, los componentes reutilizables se han limitado en general a los módulos de realización de superficie (gramáticas de generación) y la mayor parte del software disponible es para inglés.

Aparte de los sistemas experimentales que desarrollan los grupos de investigación, hay algunas empresas que ofrecen este tipo de servicios. Entre ellas, Daedalus e Inbenta, y algunas empresas internacionales con una presencia significativa en el mercado español, como Q-go y Artificial Solutions.

## Programas Educativos

La tecnología lingüística es un campo altamente interdisciplinario, que involucra la experiencia de lingüistas, informáticos, matemáticos, filósofos, psicolingüistas, y neurólogos, entre otros. En consecuencia, la formación básica actual de un lingüística computacional puede ser realizada en España en el marco de una licenciatura en Filología o Lingüística, que incluyen como asignatura troncal Lingüística Computacional, o bien en el de una ingeniería informática. Entre las universidades que ofrecen la primera opción: Universitat de Barcelona, ​​Universitat Pompeu Fabra, Universitat Oberta de Catalunya y la Universidad de Vigo. Por otro lado, las principales ingenierías que ofrecen lingüística computacional como asignatura están en: Universidad Politécnica de Madrid, Universidad Carlos III, Universidad Autónoma de Madrid, Universitat d'Alacant, Universidad Nacional de Educación a Distancia, y Euskal Herriko Unibertsitatea. Otros casos, tales como la Universidad Complutense combina ambas opciones.

Los cursos de posgrado ofrecen una formación profesional más específica. Hay varios programas de doctorado que ofrecen másters o asignaturas relacionados con el lenguaje y el procesamiento del habla. Algunas universidades como la Universitat Politècnica de Catalunya también participan en el Máster Europeo en Lengua y Habla patrocinado por ELSNET (Red Europea de Excelencia en tecnologías del lenguaje humano). Los másters a menudo se ofrecen de forma colectiva, por un grupo de universidades, ya sea a nivel estatal o a nivel europeo. Por ejemplo, la Universitat Autònoma de Barcelona ofrece el Máster Internacional en Procesamiento del Lenguaje Natural y Tecnologías del Lenguaje Humano, en colaboración con universidades extranjeras. También se ofrecen módulos de Tecnología Lingüística a los estudiantes de otros masters o cursos de doctorado, sobre todo en traducción (por ejemplo, en la Autònoma de Barcelona, ​​Alacant, Castelló, Politécnica de Valencia y Granada).

Hay 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. La Sociedad Española para el Procesamiento del Lenguaje Natural, es una organización sin ánimo de lucro con más de 300 miembros, 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.

Hay más de 30 grupos de investigación en España que trabajan en el reconocimiento del habla, el procesamiento del lenguaje natural, la traducción de texto a texto y la síntesis de voz.

La SEPLN organiza una conferencia anual a la que asisten un número creciente de investigadores que trabajan en el PLN, tanto en España como en el extranjero. La asociación edita una revista periódica y mantiene un servidor web con información sobre temas relacionados con el procesamiento del lenguaje natural y un foro abierto para los miembros de la asociación.

## Proyectos e iniciativas nacionales

El Ministerio de Educación, a través de la Comisión Interministerial de Ciencia y Tecnología (CICYT), que es el órgano que coordina y supervisa los planes estratégicos nacionales para la Ciencia y la Tecnología en España, apoya la investigación en el campo de las tecnologías de la información, a través de programas nacionales de investigación. Estos programas han impulsado numerosos proyectos de investigación y de colaboración con empresas y centros internacionales de investigación. La base del desarrollo tecnológico y de las aplicaciones comerciales para el procesamiento automático de la lengua española ha surgido en parte como resultado de estos proyectos.

El Centro para el Desarrollo Tecnológico Industrial (CDTI) es una organización pública española, dependiente del Ministerio de Ciencia e Innovación, cuyo objetivo es ayudar a las empresas españolas a incrementar su perfil tecnológico. El CDTI evalúa y financia proyectos I + D a través de programas como CENIT, AVANZA e INNPACTO.

El programa CENIT (Consorcios Estratégicos Nacionales para la Investigación Tecnológica) tiene como objetivo estimular la cooperación en I + D entre el sector privado, las universidades, los organismos y centros públicos de investigación, los parques científicos y los centros tecnológicos, impulsando la cooperación en I + D entre los sectores público y privado. Los proyectos CENIT tienen una duración mínima de cuatro años, un presupuesto mínimo de 5 millones de euros por año, y reciben una financiación mínima del 50% del sector privado. Al menos el 50% de la financiación pública se destina a centros públicos de investigación o a centros tecnológicos. Las tecnologías de la Información y la Comunicación son una de las áreas prioritarias del programa. Los proyectos en este área incluyen a menudo algún tipo de investigación en tecnologías del lenguaje.

El objetivo de los planes AVANZ@ e INNPACTO es acercar la Sociedad de la Información a los ciudadanos comunes, y a los sectores público y privado. Promover el uso de las TIC ha de tener un efecto en cadena sobre el conjunto del sector en España, y por consiguiente, en su carácter innovador. Los objetivos del Plan incluyen el incremento del porcentaje de empresas que utilizan el comercio electrónico; la promoción del uso de la facturación electrónica; la implementación de un carnet de identidad electrónico y la extensión de los trámites públicos electrónicos; alcanzar la tasa de un ordenador conectado a Internet por cada dos alumnos en las escuelas; y duplicar el número de hogares con acceso a Internet. Entre sus mayores prioridades está el facilitar el uso de las nuevas tecnologías a las personas mayores y a las personas con discapacidad, como un medio ideal para lograr la integración social, evitar la exclusión y mejorar su calidad de vida. Las herramientas basadas en tecnología lingüística constituyen el medio principal para satisfacer este objetivo (por ejemplo, los asistentes de lectura para ciegos que utilizan síntesis de voz).

Proyectos y programas anteriores han contribuido al desarrollo de una serie de herramientas de tecnología lingüística y de recursos para el español. En la siguiente sección, se resume el estado actual del soporte a la tecnología lingüística para el español.

## 

## Disponibilidad de herramientas y recursos

La siguiente tabla resume el estado actual del soporte a la tecnología lingüística para el español. La evaluación de los instrumentos y los recursos existentes se basa en estimaciones proporcionadas por un grupo de expertos en cada área.

|  | **Quantity** | **Availability** | **Quality** | **Coverage** | **Maturity** | **Sustainability** | **Adaptability** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Language Technology (Tools, Technologies and Applications)** | | | | | | | |
| Speech Recognition | 4 | 2 | 3’6 | 4’8 | 4 | 4 | 3 |
| Speech Synthesis | 4 | 2 | 4’8 | 4’8 | 4 | 4 | 3 |
| Text analysis | 3’5 | 3 | 5’4 | 5’4 | 4’5 | 3’5 | 3’5 |
| Text interpretation | 1’5 | 2 | 2’4 | 2’4 | 2 | 2 | 1 |
| Language generation | 1 | 2 | 2’4 | 2’4 | 2 | 1 | 2 |
| Machine translation | 5 | 4 | 6 | 4’8 | 5 | 3 | 3 |
| **Language Resources (Resources, Data and Knowledge Bases)** | | | | | | | |
| Text corpora | 3 | 2’5 | 3’6 | 3’6 | 3 | 3’5 | 2,5 |
| Speech corpora | 3 | 1 | 3’6 | 2’4 | 3 | 4 | 3 |
| Parallel corpora | 4 | 3 | 4’8 | 3’6 | 3 | 2 | 2 |
| Lexical resources | 3 | 3 | 3’6 | 3’6 | 3,5 | 3,5 | 2,5 |
| Grammars | 2 | 3 | 3’6 | 3’6 | 4 | 2 | 3 |

Los datos contenidos en esta tabla se pueden interpretar de forma resumida, de la siguiente manera:

* El procesamiento del habla aparece como una tecnología ligeramente más madura que el procesamiento del texto escrito. De hecho, esta tecnología ya ha sido integrada con éxito en muchas aplicaciones cotidianas, como, por ejemplo, los sistemas de diálogo hablado y las interfaces de voz para móviles y navegadores para automóviles.
* La investigación realizada hasta la fecha, ha conducido con éxito al diseño de software de calidad media-alta para el análisis de textos básicos, tales como herramientas de análisis morfológicos y de análisis sintáctico. Sin embargo, las tecnologías que requieren un procesamiento lingüístico profundo y un conocimiento semántico, son todavía muy incipientes.
* En cuanto a los recursos, existe un corpus textual de referencia de gran tamaño para el español, que contiene una mezcla equilibrada de géneros, pero que no es de fácil acceso para la investigación. Existen también diversos corpus anotados con información sintáctica, semántica y de discurso, pero no son suficientes, ni en riqueza de anotaciones ni en tamaño, para satisfacer la creciente necesidad de información lingüística.
* En particular, hay una carencia de corpus paralelos que constituyen la base de los sistemas de traducción automática estadísticos e híbridos. Existen corpus paralelos entre el español y el inglés, así como entre el español y el resto de lenguas oficiales en España. Sin embargo, faltan corpus paralelos entre el español y otros idiomas.
* Muchas de estas herramientas, recursos y formatos de codificación no se ajustan a los estándares del sector y no se pueden mantener de forma eficaz. Se requiere un plan concertado para estandarizar las interfaces de las aplicaciones y los formatos de los datos.
* Existe una situación legal confusa que restringe la utilización de textos digitales, como las publicaciones en línea, para su uso en investigación, por ejemplo para entrenar modelos estadísticos de la lengua. Los investigadores, junto con los políticos, deben intentar establecer leyes o regulaciones que les permitan utilizar para la investigación, textos a disposición del público.
* Debería intensificarse la cooperación entre la comunidad dedicada a las tecnologías lingüísticas y las relacionadas con la Web Semántica y el movimiento Linked Open Data, con objeto de establecer una base de conocimientos digitales mantenida de forma colaborativa, que pueda ser utilizada en los sistemas de información basados en la web y como base de conocimiento semántico para las aplicaciones de tecnología lingüística. Lo ideal sería que este esfuerzo fuera abordado de forma multilingüe a escala europea.

Para concluir, en una serie de áreas específicas de la investigación lingüística en español, contamos ya con un software disponible, de funcionalidad limitada. Obviamente, se requiere un mayor esfuerzo de investigación para satisfacer el déficit actual en el procesamiento de textos a un nivel semántico más profundo y para hacer frente a la falta de recursos lingüísticos, tales como corpus paralelos imprescindibles para la traducción automática.

## Comparación entre lenguas

La situación con respecto al soporte tecnológico existente para cada lengua varía considerablemente de una lengua a otra. Con el fin de comparar la situación de las distintas lenguas, esta sección presentará una evaluación basada en dos áreas de aplicación como ejemplo (traducción automática y procesamiento de voz) y una tecnología básica (análisis de textos), así como la situación de los recursos lingüísticos básicos necesarios para construir aplicaciones de tecnología lingüística.

(updated figure missing)

Figura 1: Grupos de lenguas para Procesamiento de voz

(updated figure missing)

Figura 2: Grupos de lenguas para Traducción automática

Descripción de los grupos (para el procesamiento del habla y traducción automática)

* Grupo 1 (excelente soporte tecnológico): Dispone de tecnologías de excelente calidad y rendimiento que pueden utilizarse prácticamente en todas las aplicaciones necesarias.
* Grupo 2 (buen soporte): Dispone de tecnologías de calidad y rendimiento razonables cuya usabilidad está limitada a ciertas aplicaciones o dominios.
* Grupo 3 (apoyo medio): Dispone de prototipos de investigación, aplicaciones comerciales iniciales o servicios gratuitos con diferentes grados de calidad y rendimiento.
* Grupo 4 (de menos a casi ningún apoyo): Dispone de prototipos rudimentarios con una calidad y rendimiento muy limitados.

(updated figure missing)

Figura 3: Grupos de lenguas para Análisis de texto

(updated figure missing)

Figura 4: Grupos de lenguas para Recursos lingüísticos

Descripción de los grupos (para el análisis de texto y recursos lingüísticos)

* Grupo 1 (excelente soporte tecnológico): Dispone de tecnologías y recursos de uso extensivo, que cubren prácticamente todos los fenómenos lingüísticos relevantes para la lengua.
* Grupo 2 (muy buen soporte): Dispone de tecnologías y recursos que se están usando en una diversidad de aplicaciones y que cubren los fenómenos lingüísticos más importantes.
* Grupo 3 (buen soporte): Dispone de tecnologías y recursos que cubren un porcentaje razonable de fenómenos lingüísticos y que se utilizan en aplicaciones de dominio restringido.
* Grupo 4 (apoyo medio): Dispone de prototipos y recursos de investigación, con diferentes grados de calidad y cubertura.
* Grupo 5 (de menos a casi ningún apoyo): Dispone de prototipos rudimentarios con una calidad y cubertura muy limitados.

Las tablas anteriores muestran que, gracias a la financiación a gran escala de las últimas décadas en el área de las tecnologías lingüísticas, la lengua española está mejor equipada que otros idiomas. Su situación es comparable a la de los otros idiomas grandes, como el francés y el alemán. Sin embargo, los recursos y las herramientas disponibles no alcanzan todavía la cobertura y la calidad de los que existen para el inglés, que está a la cabeza en todas las áreas. Y hay que tener en cuenta que incluso para esta lengua los recursos y herramientas presentan deficiencias importantes con respecto a las aplicaciones de alta calidad.

En el caso del procesamiento del habla para el español, la tecnología actual ya es capaz de integrarse con éxito en una serie de aplicaciones industriales, como por ejemplo el diálogo hablado y los sistemas de dictado. Los módulos de análisis de texto y los recursos lingüísticos existentes cubren ya la mayoría de fenómenos morfosintácticos del español y forman parte de muchas aplicaciones que implican un procesamiento superficial del lenguaje natural, como por ejemplo, la corrección ortográfica y la ayuda a la edición.

Sin embargo, para la construcción de aplicaciones más sofisticadas, como por ejemplo, de traducción automática, existe una necesidad acuciante de recursos y tecnologías que cubran una gama más amplia de aspectos lingüísticos y permitan un análisis semántico en profundidad del texto de entrada. Mediante la mejora de la calidad y cobertura de estos recursos y tecnologías básicas, seremos capaces de abordar una gama más amplia de áreas de aplicación avanzada, incluyendo traducción automática de alta calidad.

## Conclusiones

**En esta serie de libros blancos, se ha llevado a cabo un importante esfuerzo inicial con objeto de evaluar el soporte tecnológico de 30 lenguas europeas. Al identificar las carencias, necesidades y deficiencias, las partes interesadas ​​están ahora en condiciones de diseñar un programa de investigación y desarrollo a gran escala dirigido a la construcción de una Europa verdaderamente multilingüe y tecnológicamente capaz.**

Hemos visto que existen grandes diferencias entre las lenguas de Europa. Si bien algunas lenguas cuentan con aplicaciones y recursos de calidad para algunas áreas concretas de aplicación, las demás (por lo general lenguas "más pequeñas") adolecen de importantes lagunas. Muchas lenguas carecen de las tecnologías básicas para el análisis textual y de los recursos esenciales para el desarrollo de estas tecnologías. Otras cuentan con herramientas y recursos básicos, pero son todavía incapaces de invertir recursos en el procesamiento semántico. Por lo tanto, todavía debemos hacer un esfuerzo a gran escala para alcanzar el ambicioso objetivo de proporcionar traducción automática de alta calidad entre todos los idiomas europeos.

Podemos ser moderadamente optimistas acerca del apoyo tecnológico a la lengua española. En España existe una pequeña industria lingüística y un marco de investigación que se benefició en el pasado de programas de investigación importantes. Se han producido y distribuido una serie de recursos y de tecnologías de última generación para el español. Sin embargo, tanto el tamaño de los recursos como el número de herramientas son todavía muy limitados en comparación con los recursos y las herramientas existentes para el inglés, y, desde luego, no son lo suficientemente completos como para dar el apoyo tecnológico integral que necesita una sociedad del conocimiento verdaderamente multilingüe.

Desgraciadamente, la implicación de la industria en las tecnologías lingüísticas para el español en la actualidad es reducida. La mayoría de grandes empresas han interrumpido o reducido mucho sus actividades en este área, dejándola mayoritariamente en manos un pequeño grupo de empresas medianas y pequeñas, más especializadas, que no pueden afrontar un mercado internacional en el que la barrera del idioma es un factor clave que frena el comercio electrónico transfronterizo en la UE .[[20]](#endnote-20)

Está claro que debe hacerse un mayor esfuerzo para crear recursos lingüísticos para el español, así como impulsar la investigación, la innovación y el desarrollo en general. La necesidad de grandes cantidades de datos y la gran complejidad de las aplicaciones tecnológicas lingüísticas hace que sea vital el desarrollo de una nueva infraestructura para estimular un mayor intercambio y cooperación.

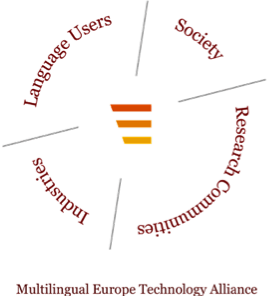
También se detecta una falta de continuidad en la financiación de la investigación y el desarrollo. Suelen alternarse programas coordinados a corto plazo, con períodos de escasa o nula financiación. Además, hay una falta general de coordinación con los programas de otros países de la UE y a nivel de la Comisión Europea.

Un gran esfuerzo coordinado centrado en las tecnologías lingüísticas ayudaría a la lengua española, y a los otros idiomas europeos, y contribuiría a establecer una verdadera agenda multilingüe para Europa y para el mundo en su conjunto.[[21]](#endnote-21)

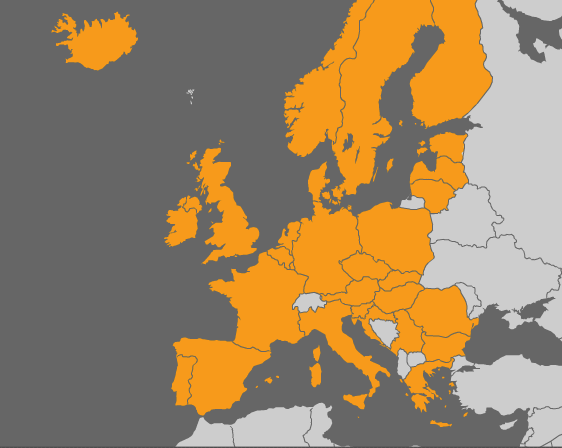
El objetivo a largo plazo de META-NET es impulsar la tecnología lingüística de alta calidad para todas las lenguas, a fin de lograr la unidad política y económica a través de la diversidad cultural. La tecnología ayudará a derribar las barreras existentes y a construir puentes entre las lenguas de Europa. Esto requiere que todas las partes interesadas - en el ámbito político, la investigación, la empresa y la sociedad – unan sus esfuerzos de cara al futuro.[[22]](#endnote-22)

# Acerca de META-NET

META-NET es una Red de Excelencia financiada por la Comisión Europea. La red consta actualmente de 47 miembros de 31 países europeos. META-NET fomenta la Alianza Tecnológica Multilingüe de Europa, una amplia comunidad europea de profesionales y organizaciones vinculadas con las tecnologías lingüísticas.



Alianza Tecnológica Multilingüe de Europa (META)



Países representados en META-NET

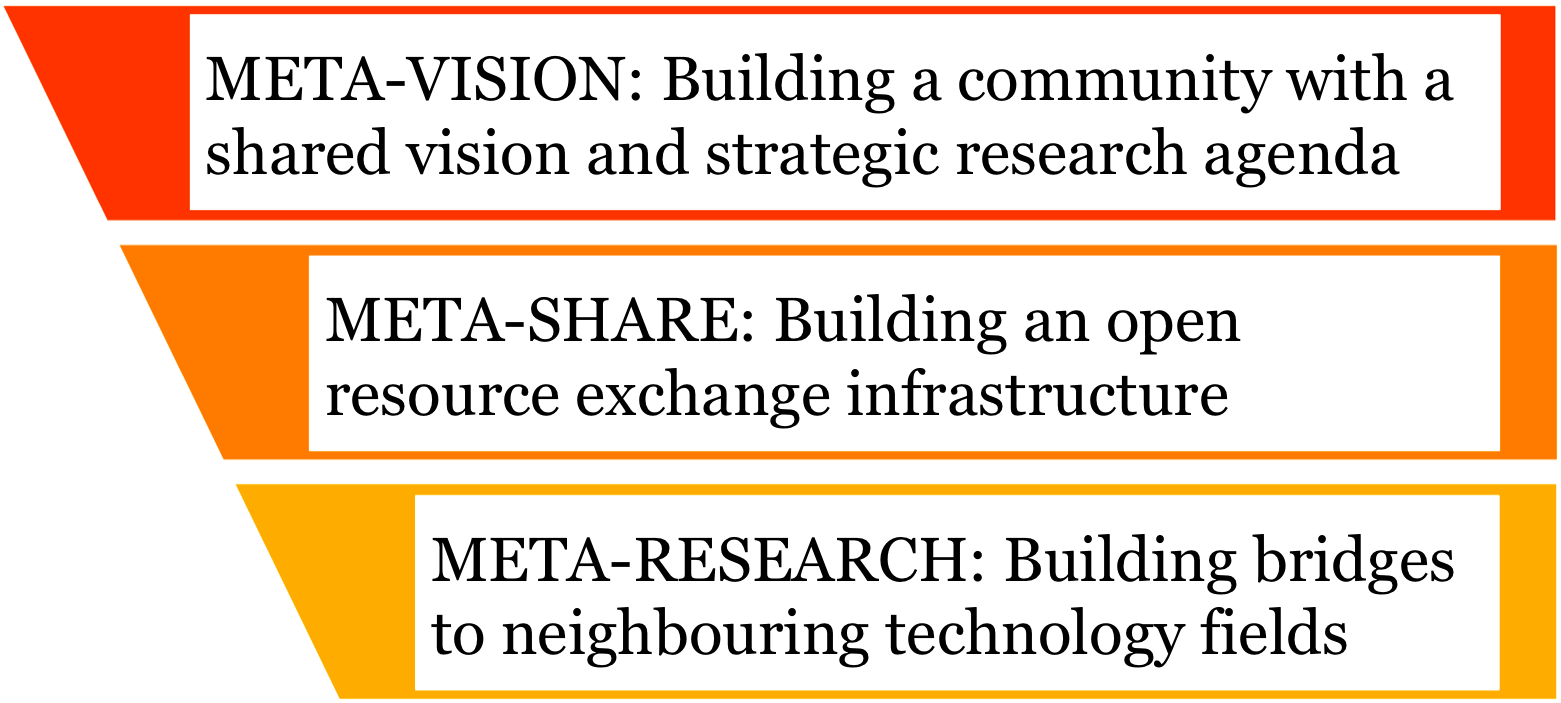
META-NET promueve los fundamentos tecnológicos necesarios para una sociedad de la información europea verdaderamente multilingüe que:

* haga posible la comunicación y la cooperación entre las distintas lenguas;
* proporcione un acceso igualitario a la información y al conocimiento en cualquier lengua;
* ofrezca tecnología de la información en red, avanzada y asequible, a todos los ciudadanos europeos.

La red apoya una Europa unida con un mercado digital y un espacio de información únicos. Estimula y fomenta tecnologías multilingües para todas las lenguas europeas. Estas tecnologías están en la base de la traducción automática, la producción de contenidos, el procesamiento de la información y la gestión del conocimiento para una gran variedad de aplicaciones y ámbitos temáticos. La red pretende mejorar las estrategias actuales de forma que se favorezcan la comunicación y la cooperación entre las distintas lenguas. Los europeos tienen derecho al mismo nivel de información y conocimiento independientemente de cuál sea su lengua de origen.

## Lineas de Acción

META-NET surgió el 1 de febrero de 2010 con el objetivo de potenciar la investigación en tecnologías de la lengua. META-NET ha llevado a cabo varias actividades para promover sus objetivos. META-VISIÓN, META-SHARE y META-RESEARCH son las tres líneas de acción de la red.



Las tres lineas de acción en META-NET

**META-VISION** fomenta una comunidad ​​dinámica e influyente, formada por todas las partes interesadas, que se une en torno a una visión compartida y una agenda común estratégica de investigación. El objetivo principal de esta actividad es crear una comunidad europea de tecnología lingüística, coherente y cohesionada poniendo en contacto a representantes de un campo muy diverso y fragmentado. El presente Libro Blanco ha sido preparado de forma conjunta para 30 idiomas. La visión tecnológica compartida se desarrolló en tres grupos de trabajo distribuidos por ámbitos. Se creó el Consejo Tecnológico META para discutir y preparar la agenda estratégica sobre la base del resultado de estos grupos en estrecha interacción con toda la comunidad de tecnologías lingüísticas.

**META-SHARE** constituye una instalación abierta y distribuida para intercambiar y compartir recursos. La red “punto a punto” (P2P) de repositorios contendrá datos lingüísticos, herramientas y servicios web documentados con metadatos de alta calidad y organizados en categorías estandarizadas. El acceso y la búsqueda de los recursos resulta así fácil y uniforme. Los recursos disponibles incluyen materiales gratuitos y de código abierto, así como artículos comercializables y sujetos a cuotas.

**META-RESEARCH** tiende puentes hacia ámbitos tecnológicos cercanos. Esta actividad tiene por objeto utilizar los avances en otros campos y sacar provecho de investigaciones innovadoras que puedan beneficiar a la tecnología de la lengua. En particular, esta actividad pretende enriquecer la investigación puntera en traducción automática recogiendo y preparando datos y organizando recursos lingüísticos para la evaluación de algoritmos y tecnologías; elaborando inventarios de herramientas y metodologías; y organizando talleres y eventos formativos para los miembros de la comunidad.

## Organizaciones miembros

La siguiente tabla muestra las organizaciones que participan en META-NET y sus representantes.

| País | Organización | Participantes |
| --- | --- | --- |
| Alemania | DFKI | Hans Uszkoreit y Georg Rehm |
|  | Universidad Técnica de Aquisgrán | Hermann Ney |
|  | Universidad del Sarre | Manfred Pinkal |
| Austria | Universidad de Viena | Gerhard Budin |
| Bélgica | Universidad de Amberes | Walter Daelemans |
|  | Universidad de Lovaina | Dirk van Compernolle |
| Bulgaria | Academia Búlgara de Ciencias | Svetla Koeva |
| Chipre | Universidad de Chipre | Jack Burston |
| Croacia | Universidad de Zagreb | Marko Tadić |
| Dinamarca | Universidad de Copenhague | Bolette Sandford Pedersen y Bente Maegaard |
| Eslovaquia | Academia Eslovaca de Ciencias | Radovan Garabik |
| Eslovenia | Instituto Jozef Stefan | Marko Grobelnik |
| España | Barcelona Media – Centro de Innovación | Toni Badia y Maite Melero |
|  | Universitat Politècnica de Catalunya | Asunción Moreno |
|  | Universitat Pompeu Fabra | Núria Bel |
| Estonia | Universidad de Tartu | Tiit Roosmaa |
| Finlandia | Universidad Aalto | Timo Honkela |
|  | Universidad de Helsinki | Kimmo Koskenniemi y Krister Linden |
| Francia | CNRS/LIMSI | Joseph Mariani |
|  | Agencia de Distribución de Recursos Lingüísticos y de Evaluación (ELRA) | Khalid Choukri |
| Grecia | Instituto para el Procesamiento de Voz y Lenguaje, “Athena” R.C. | Stelios Piperidis |
| Hungría | Academia Húngara de Ciencias | Tamás Váradi |
|  | Universidad de Tecnología y Economía de Budapest | Géza Németh and Gábor Olaszy |
| Irlanda | Dublin City University | Josef van Genabith |
| Islandia | Universidad de Islandia | Eirikur Rögnvaldsson |
| Italia | Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale “Antonio Zampolli” | Nicoletta Calzolari |
|  | Fondazione Bruno Kessler | Bernardo Magnini |
| Letonia | Tilde | Andrejs Vasiljevs |
|  | Instituto de Matemáticas e Informática, Universidad de Letonia | Inguna Skadina |
| Lituania | Instituto de la Lengua Lituana | Jolanta Zabarskaitė |
| Luxemburgo | Arax Ltd. | Vartkes Goetcherian |
| Malta | Universidad de Malta | Mike Rosner |
| Noruega | Universidad de Bergen | Koenraad De Smedt |
| Países Bajos | Universidad de Utrecht | Jan Odijk |
|  | Universidad de Groningen | Gertjan van Noord |
| Polonia | Academia Polaca de Ciencias | Adam Przepiórkowski y Maciej Ogrodniczuk |
|  | Universidad de Lodz | Barbara Lewandowska-Tomaszczyk y Piotr Pęzik |
| Portugal | Universidad de Lisboa | Antonio Branco |
|  | Instituto de Ingeniería de Sistemas y Computación | Isabel Trancoso |
| Reino Unido | Universidad de Manchester | Sophia Ananiadou |
|  | Universidad de Edimburgo | Steve Renals |
| República Checa | Universidad Charles de Praga | Jan Hajic |
| Rumanía | Academia Rumana de Ciencias | Dan Tufis |
|  | Universidad Alexandru Ioan Cuza | Dan Cristea |
| Serbia | Universidad de Belgrado | Dusko Vitas, Cvetana Krstev y Ivan Obradovic |
|  | Instituto Mihailo Pupin | Sanja Vranes |
| Suecia | Universidad de Gothenburg | Lars Borin |

## Agradecimientos

Dr. Toni Badia, Universitat Pompeu Fabra and Barcelona Media Centre d‘Innovació

Dra. Núria Bel, Universitat Pompeu Fabra

Dr. Aljoscha Burchardt, DFKI

Kathrin Eichler, DFKI

Dra. Mª Antònia Martí, Universitat de Barcelona

Dra. Asunción Moreno, Universitat Politècnica de Catalunya.

Dr. Lluís Padró, Universitat Politècnica de Catalunya

Dr. GeorgRehm, DFKI

# Referencias

1. European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public\_opinion/flash/fl\_313\_en.pdf). [↑](#endnote-ref-1)
2. European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008\_0566\_en.pdf). [↑](#endnote-ref-2)
3. UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf). [↑](#endnote-ref-3)
4. European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies). [↑](#endnote-ref-4)
5. <http://www.ethnologue.org> [↑](#endnote-ref-5)
6. Moreno Fernández, Francisco y Otero Roth, Jaime (2007) Atlas de la lengua española en el mundo. Editorial Planeta. [↑](#endnote-ref-6)
7. -Nueva gramática de la lengua española (2009), Real Academia Española, Colección Nebrija y Bello, Espasa. [↑](#endnote-ref-7)
8. <http://www.fundeu.es/> [↑](#endnote-ref-8)
9. <http://www.wikilengua.org/index.php/Portada> [↑](#endnote-ref-9)
10. <http://www.efnil.org/documents/language-legislation-version-2007/spain/spain> [↑](#endnote-ref-10)
11. -Vivanco Torres, Hiram. Enseñanza del español como lengua extranjera. Ponencia Congreso Internacional de la Lengua Española. Cartagena. 2007. (<http://congresosdelalengua.es/cartagena/ponencias/seccion_3/36/vivanco_hiram.htm>) [↑](#endnote-ref-11)
12. Garcia Delgado, Jose Luis (2007) Economia del español, una introducción. Editorial Planeta. [↑](#endnote-ref-12)
13. <http://www.fundacionlengua.com/es/valor-economico-espanol-genera-35-millones-empleos/art/1599/> [↑](#endnote-ref-13)
14. Rojo, Guillermo y Sánchez, Mercedes (2010) El español en la red. Editorial Ariel. (<http://www.fundacion.telefonica.com/es/debateyconocimiento/media/publicaciones/espanol_red.pdf>) [↑](#endnote-ref-14)
15. [http://www.britishcouncil.org/learning-research-english-next.pdf](%20http:/www.britishcouncil.org/learning-research-english-next.pdf) [↑](#endnote-ref-15)
16. [http://www.internetworldstats.com/stats7.htm](%20http:/www.internetworldstats.com/stats7.htm) [↑](#endnote-ref-16)
17. [http://www.nielsen-online.com/intlpage.html](%20http:/www.nielsen-online.com/intlpage.html) [↑](#endnote-ref-17)
18. <http://www.ontsi.red.es/> [↑](#endnote-ref-18)
19. Quirós Romero, Cipriano (2010) Lengua y tecnologías de la información y las comunicaciones. Editorial Planeta. [↑](#endnote-ref-19)
20. <http://ec.europa.eu/consumers/strategy/docs/com_staff_wp2009_en.pdf> [↑](#endnote-ref-20)
21. V. Reding, J.Figel’, Preface, in Human Language Technologies for Europe, TC-Star project, http://www.tcstar.org/pubblicazioni/D17\_HLT\_ENG.pdf [↑](#endnote-ref-21)
22. V. Reding, J.Figel’, Preface, in *Human Language Technologies for Europe*, TC-Star project, <http://www.tcstar.org/pubblicazioni/D17_HLT_ENG.pdf> [↑](#endnote-ref-22)