META NET

White Paper Series Serie de Libros Blancos

THE SPANISH LANGUAGE IN THE DIGITAL AGE

LA LENGUA ESPAÑOLA EN LA ERA DIGITAL

Maite Melero
Toni Badia
Asunción Moreno

Springer

White Paper Series Serie de Libros Blancos

THE SPANISH LANGUAGE IN THE DIGITAL AGE

LA LENGUA ESPAÑOLA EN LA ERA DIGITAL

Maite Melero Barcelona Media Centre d'Innovació
Toni Badia Universitat Pompeu Fabra
Asunción Moreno Universitat Politècnica de Catalunya

Georg Rehm, Hans Uszkoreit
(editores, editors)

Editors
Georg Rehm
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: georg.rehm@dfki.de

Hans Uszkoreit
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: hans.uszkoreit@dfki.de

ISSN 2194-1416 ISSN 2194-1424 (electronic)
ISBN 978-3-642-30840-6 ISBN 978-3-642-30841-3 (eBook)
DOI 10.1007/978-3-642-30841-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012946618

PRÓLOGO PREFACE

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

La disponibilidad y el uso de las tecnologías lingüísticas en Euro-pa varían según el idioma. Por lo tanto, las acciones requeridas para apoyar la investigación y el desarrollo de estas tecnologías también varían para cada idioma. Las acciones necesarias también dependen de factores como la complejidad intrínseca de la lengua 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 para cada lengua (p. 79). 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 sugieren que existen todavía muchas lagunas por cubrir en este área.

META-NET se compone de 54 centros de investigación de 33 países que están trabajando con representantes de empresas comerciales, agencias gubernamentales, industria, organizaciones de investigación, empresas de software, proveedores de tecnología y universidades europeas [1].

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.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others. The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series (p. 79). The analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of additional research.

As of November 2011, META-NET consists of 54 research centres from 33 European countries [1]. META-NET is working with stakeholders from economy (software companies, technology providers and users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

Los autores de este documento agradecen a los autores del Libro Blanco para el alemán el permiso para reutilizar material seleccionado de su texto original [2].

La elaboración de este Libro Blanco ha sido financiada por el Séptimo Programa Marco y el Programa de Políticas de Apoyo a las TIC de la Comisión Europea mediante los contratos T4ME (GA 249 119), CESAR (GA 271 022), META-NET4U (GA 270 893) i META-NORD (GA 270 899).

The authors of this document are grateful to the authors of the White Paper on German for permission to re-use selected language-independent materials from their document [2].

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249 119), CESAR (Grant Agreement 271 022), METANET4U (Grant Agreement 270 893) and META-NORD (Grant Agreement 270 899).

CONTENIDO CONTENTS

LA LENGUA ESPAÑOLA EN LA ERA DIGITAL

THE SPANISH LANGUAGE IN THE DIGITAL AGE

1

RESUMEN EJECUTIVO

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.

La tecnología lingüística construye puentes para el futuro de Europa.

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.

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 resul-

tados 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.

Las tecnologías de la lengua, clave para el futuro.

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.

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.

Las tecnologías lingüísticas contribuyen a unificar Europa.