META NET

White Paper Series Serie de Libros Brancos

THE GALICIAN LANGUAGE IN THE DIGITAL AGE

O IDIOMA GALEGO NA ERA DIXITAL

Carmen García Mateo
Montserrat Arza Rodríguez

Springer

White Paper Series Serie de Libros Brancos

THE GALICIAN LANGUAGE IN THE DIGITAL AGE

O IDIOMA GALEGO NA ERA DIXITAL

Carmen García Mateo Univ. de Vigo
Montserrat Arza Rodríguez Univ. de Vigo

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 978-3-642-30799-7 ISBN 978-3-642-30798-0 (eBook)
DOI 10.1007/978-3-642-30799-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012946622

PREFACIO PREFACE

Este libro branco é parte dunha serie de libros que fomenta o coñecemento sobre a tecnoloxía lingüística (TL) e o seu potencial. Está dirixida a xornalistas, políticos, comunidades lingüísticas, profesores de idiomas, e público en xeral.

A cobertura e o uso de tecnoloxías lingüísticas en Europa varía dun idioma a outro. Como consecuencia, as accións precisas para dar apoio á investigación e o desenvolvemento varían, e os pasos a seguir dependen de diversos factores, tales como a complexidade da lingua ou a dimensión da súa comunidade.

META-NET, unha Rede de Excelencia da Comisión Europea, afrontou este reto poñendo en marcha unha análise da situación actual das tecnoloxías e dos recursos lingüísticos (p. 77). A análise céntrase en 23 linguas europeas oficiais e en varias linguas rexionais de relevancia. Os resultados da análise suxiren que en cada lingua existen moitas carencias significativas. A análise e avaliación detalladas da situación de cada unha das linguas por parte de expertos axudará a maximizar o impacto das tecnoloxías lingüísticas e a minimizar calquera risco asociado.

Con data Novembro de 2011, META-NET está formada por 54 centros de investigación de 33 países europeos (p. 73). META-NET está a traballar con partes interesadas de economía (compañías de software, provedores de tecnoloxía e usuarios), axencias gobernamentais, organismos de investigación, organizacións non-gobernamentais, asociacións e universidades europeas. Xuntamente con elas, META-NET está a crear unha visión da tecnoloxía común e unha axenda de investigación estratéxica para a Europa multilingüe do 2020.

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. 77). 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 (p. 73). 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.

As autoras deste documento agradecen aos autores do "Libro Branco sobre o alemán" [1] o seu consentimento para reutilizar material seleccionado do seu documento orixinal. Asemesmo, as autoras agradecen a colaboración dos seguintes expertos no idioma galego: Dr. Xavier G. Guinovart (Universidade de Vigo), Dr. Eduardo R. Banga (Universidade de Vigo), Dr. Xosé Luis Regueira (Universidade de Santiago de Compostela), e Don José Ramom Pichel (Imaxin Software). Iinformación das páxinas web do Consello da Cultura Galega (Proxecto LOIA) e da Secretaría Xeral de Política Lingüística – Xunta de Galicia foi empregada na elaboración deste texto.

O desenvolvemento deste libro branco foi financiado polo Sétimo Programa Marco e o Programa de apoio ás TIC (ICT Policy support programme) da Comisión Europea en virtude dos contratos T4ME (acordo de subvención 249 119), CESAR (acordo de subvención 271 022), METANET4U (acordo de subvención 270 893) e META-NORD (acordo de subvención 270 899).

The authors of this document are grateful to the authors of the White Paper on German [1] for permission to re-use selected language-independent materials from their document. Furthermore, the authors would like to thank Dr. Xavier G. Guinovart (University of Vigo), Dr. Eduardo R. Banga (University of Vigo), Dr. Xosé Luis Regueira (University of Santiago de Compostela) and Mr. José Ramom Pichel (Imaxin Software) for their contributions to this white paper. Material availabe on the web sites of "Consello da Cultura Galega – Proxecto LOIA" and "Secretaría Xeral de Política Lingüística – Xunta de Galicia" has been used.

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).

ÍNDICE CONTENTS

O IDIOMA GALEGO NA ERA DIXITAL

THE GALICIAN LANGUAGE IN THE DIGITAL AGE

RESUMO

A lingua é o modo esencial de comunicación entre humanos. A lingua permítenos expresar ideas e sentimentos, axúdanos a aprender e ensinar, é esencial para vivir, é o vehículo preferido para a transmisión de cultura, e é un símbolo de identidade.

Co noso nivel actual de globalización, temos moitas formas de comunicarnos facilmente con xente de todo o mundo. Por exemplo, as novas tecnoloxías da información e as comunicacións facilitaron o desenvolvemento das redes sociais que impulsan e realzan a interacción entre xente de practicamente todos os países e culturas. Tamén, nos últimos anos, asistimos a fluxos importantes de xente estranxeira entre os nosos países, p. ex. o turismo ou a inmigración, que crean a necesidade de comunicación entre linguas diferentes. Este problema de comunicación interidiomático é a miúdo resolto mediante o uso dunha lingua franca.

Os países de Europa son un exemplo claro de diversidade lingüística e cultural a pesar do feito de que, durante os últimos 60 anos, Europa converteuse nunha estrutura política e económica ben definida. Isto significa que a comunicación entre cidadáns europeos, tanto a máis cotiá como a que se produce no eido dos negocios ou a política, ben sexa en galego como en grego, en italiano ou en islandés, vese confrontada inevitablemente polas barreiras do idioma. As institucións da UE gastan preto de mil millóns de euros ao ano para manter a súa política de plurilingüismo, é dicir, na tradución de textos e a interpretación de comunicacións orais. Paralelamente, o inglés está a converterse nunha lingua franca na comunicación entre os cidadáns europeos.

En España, atopamos un escenario similar. España ten unha lingua oficial, o español, tamén coñecido como castelán, e tres linguas cooficiais: galego, catalán, e vasco. A preservación do multilingüismo en España non foi unha tarefa fácil. É o resultado dun proceso complexo para intencionalmente preservar a identidade cultural e lingüística dentro e entre as diversas rexións e habitantes de España. De forma similar ao uso do inglés no caso europeo, a comunicación directa entre cidadáns de áreas diferentes de España, a miúdo necesita utilizar o castelán como lingua franca.

A tecnoloxía da linguaxe constrúe pontes para o futuro de Europa.

En ambos os dous niveis, o europeo e o español, o multilingüismo é un patrimonio cultural que é preciso conservar. A globalización non se debería converter nun mecanismo que promova o abandono do noso rico patrimonio lingüístico e cultural, invitándonos a abandonar o uso da nosa propia lingua a favor dunha lingua franca. Nun entorno de comunicación globalizado, deberiamos atopar formas de comunicarnos co mundo ao mesmo tempo que preservamos a nosa propia lingua e, con iso, a nosa identidade cultural.

A tecnoloxía da lingua e a investigación lingüística poden facer unha contribución significativa para salvar estas barreiras lingüísticas. Combinada con dispositivos e aplicacións intelixentes, no futuro a tecnoloxía da lingua será capaz de axudar aos cidadáns a falar uns cos outros de forma sinxela, así como a traballar uns cos ou-

tros aínda que non se fale unha lingua común. De xeito que as solucións da tecnoloxía da lingua ao final servirán como unha ponte exclusiva entre linguas diferentes. Non obstante, as ferramentas de tecnoloxías da lingua e fala actualmente dispoñibles no mercado (que van de sistemas de respuesta automática a interfaces de lingua naturais – incluíndo sistemas de tradución e ferramentas de resumo, entre moitas outras), están a día de hoxe por debaixo de lograren este ambicioso obxectivo.

Xa na década de 1970, a UE decatouse da grande relevancia das tecnoloxías da linguaxe como factor impulsor da unidade europea, e comezou a financiar os seus primeiros proxectos de investigación. Ao mesmo tempo, establecéronse proxectos nacionais que xeraron valiosos resultados, pero nunca chegaron a constituír unha acción europea coordinada. Os actores dominantes neste campo son sobre todo empresas privadas con base en Norte América. Hoxe en día, os modelos predominantes nas tecnoloxías da linguaxe dependen de métodos estatísticos que non fan uso de métodos ou coñecementos lingüísticos profundos. Por exemplo, as frases tradúcense automaticamente mediante a comparación dunha nova frase con miles de frases previamente traducidas por humanos.

As tecnoloxías da linguaxe axudan a unificar Europa.

A calidade da tradución depende en gran medida da cantidade e calidade dos corpus de mostra dispoñibles. Mentres que a tradución automática de oracións simples en idiomas cunha cantidade suficiente de material textual dispoñible pode dar resultados útiles, os métodos estatísticos están condenados ao fracaso no caso de idiomas cun material de mostra moito máis pequeno, ou no caso da tradución de oracións con estruturas complexas. Analizar as propiedades estruturais máis profundas das linguas é a única forma cara adiante para construír aplicacións que funcionen ben para unha serie ampla de linguas.

A solución ao problema de comunicación entre linguas é, polo tanto, desenvolver tecnoloxías facilitadoras. Para lograr este obxectivo e conservar a diversidade cultural e lingüística de Europa, é necesario realizar primeiro unha análise sistemática das particularidades lingüísticas de todas as linguas europeas, ademais da análise do estado actual de cadansúa tecnoloxía da lingua. Este é o propósito do presente libro acerca da lingua galega.

Este volume mostra unha análise detallada das tecnoloxías da lingua, aplicacións e solucións para o galego. Atopámonos con que hai unha cantidade bastante reducida de produtos, tecnoloxías e recursos deseñados para o galego. Hai algunhas ferramentas para síntese de voz, recoñecemento de fala, corrección de ortografía e comprobación de gramática. Contamos tamén con algunhas aplicacións de tradución automática, na súa maioría entre español e galego. Como esta serie de libros brancos sobre as linguas europeas demostra, hai grandes diferenzas en canto a investimento en solucións tecnolóxicas e ao estado da investigación lingüística entre os Estados membros de Europa. O galego é unha das linguas da UE que necesita aínda máis investigación para que as solucións das tecnoloxías da linguaxe cheguen a ser verdadeiramente eficaces e poidan ser empregadas para un uso diario. Ao mesmo tempo, existen boas perspectivas para acadar unha posición salientable nesta área tecnolóxica tan importante. Este desenvolvemento de tecnoloxía da lingua de alta calidade é urxente, tendo moita importancia para a conservación dunha lingua minorizada e minoritaria como galego.