META NET

White Paper Series Coleção Livros Brancos

THE PORTUGUESE LANGUAGE IN THE DIGITAL AGE

A LÍNGUA PORTUGUESA NA ERA DIGITAL

António Branco
Amália Mendes
Sílvia Pereira
Paulo Henriques
Thomas Pellegrini
Hugo Meinedo
Isabel Trancoso
Paulo Quaresma
Vera Lúcia Strube de Lima
Fernanda Bacelar

Springer

White Paper Series

Coleção Livros Brancos

THE PORTUGUESE LANGUAGE IN THE DIGITAL AGE

A LÍNGUA PORTUGUESA NA ERA DIGITAL

António Branco Universidade de Lisboa
Amália Mendes CLUL, Universidade de Lisboa
Sílvia Pereira Universidade de Lisboa
Paulo Henriques CLUL, Universidade de Lisboa
Thomas Pellegrini INESC-ID
Hugo Meinedo INESC-ID
Isabel Trancoso INESC-ID, IST
Paulo Quaresma Universidade de Évora
Vera Lúcia Strube de Lima PUCRS
Fernanda Bacelar CLUL, Universidade de Lisboa

Georg Rehm, Hans Uszkoreit
(organizadores, 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 978-3-642-29592-8 ISBN 978-3-642-29593-5 (eBook)
DOI 10.1007/978-3-642-29593-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012939476

© Springer-Verlag Berlin Heidelberg 2012
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

PREFÁCIO PREFACE

Este Livro Branco, sobre a língua portuguesa na era digital, faz parte de uma coleção que promove o conhecimento sobre a tecnologia da linguagem e o seu potencial. É dirigido a um público o mais vasto possível, não especializado nestas matérias, incluindo comunidades linguísticas, jornalistas, políticos ou docentes, entre muitos outros.

Este livro procura disponibilizar uma análise do estado de desenvolvimento da tecnologia da linguagem para a língua portuguesa, assim como das perspectivas que se oferecem, e das ações necessárias, para a consolidação do português como língua de comunicação internacional com projeção global, no quadro desta tecnologia emergente.

Esta coleção de Livros Brancos foi organizada pela META-NET, uma Rede de Excelência parcialmente financiada pela Comissão Europeia, que levou a cabo uma análise dos recursos e tecnologias da linguagem atualmente disponíveis. A análise abordou as 23 línguas oficiais europeias assim como outras línguas importantes na Europa a nível nacional e regional.

Em Novembro de 2011, a rede META-NET integrava 54 centros de investigação de 33 países europeus (p. 81). Esta rede está a colaborar com atores do setor da economia, agências governamentais, instituições de investigação, organizações não governamentais, comunidades linguísticas e universidades. Em conjunto com todos estes atores, a META-NET procura estimular uma agenda de investigação estratégica partilhada para uma Europa e para um mundo multilingue.

This white paper about the Portuguese language in the digital age is part of a series that promotes knowledge about language technology and its potential. It addresses a wider non expert audience, in general, including language communities, journalists, politicians or educators, among many others.

This book seeks to make available an assessment of the state of development of language technology for Portuguese, and reports on the prospects, and necessary actions, for the consolidation of Portuguese as a language for international communication with global projection, in the scope of this emerging technology.

The present White Paper series was organized by META-NET, a Network of Excellence partially funded by the European Commission, which has conducted an analysis of current language resources and technology. The analysis focused on the 23 official European languages as well as other important national and regional languages in Europe.

As of November 2011, META-NET consists of 54 research centres from 33 European countries (p. 81). It is working with stakeholders from economy, government agencies, research organisations, non governmental organisations, language communities and universities. Together with all these actors, META-NET seeks to foster a shared strategic research agenda for a multilingual Europe and a multilingual world.

Os autores deste documento agradecem aos autores do Livro Branco sobre o alemão por permitirem a utilização de partes seleccionadas do seu texto original [1].

A realização deste Livro Branco foi financiada pelo 7º Programa-Quadro e pelo Programa de Apoio à Política das TIC (ICT PSP) da Comunidade Europeia no âmbito dos contratos T4ME (Acordo de Financiamento 249119), CESAR (Acordo de Financiamento 271022), METANET4U (Acordo de Financiamento 270893) e META-NORD (Acordo de Financiamento 270899).

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 [1].

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 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

ÍNDICE CONTENTS

A LÍNGUA PORTUGUESA NA ERA DIGITAL

THE PORTUGUESE LANGUAGE IN THE DIGITAL AGE

SUMÁRIO EXECUTIVO

A linguagem humana é uma porta para o mundo que nos rodeia. É usando a linguagem no dia a dia que comunicamos, aprendemos, trocamos informação, planeamos o nosso futuro, nos coordenamos uns com os outros para melhor agirmos em conjunto, efabulamos ou nos comprazemos com a leitura de uma história ou de um poema.

Porém, na era digital e num mundo globalizado, a linguagem humana é também uma das maiores barreiras comunicacionais com que nos deparamos. As novas tecnologias da informação e da comunicação colocam ao nosso alcance pessoas de todo o mundo com quem será possível interagir, assim como um acervo ilimitado de informação a que será possível aceder. No entanto, para cada um de nós, este novo universo, na sua quase totalidade, continua inacessível e fechado, encerrado nas fronteiras invísiveis das línguas que o dividem.

A Europa será talvez um caso paradigmático do impacto resultante das barreiras linguísticas. Durante os últimos 60 anos, tornou-se numa estrutura política e económica com identidade própria. Tem um imenso património quer do ponto de vista da diversidade cultural quer do ponto de vista da diversidade linguística. Contudo, da língua portuguesa à polaca ou da italiana à islandesa, os cidadãos europeus são confrontados com a dificuldade de comunicar entre si em diferentes línguas, tanto no dia a dia, como na esfera dos negócios ou da política. As instituições da União Europeia, por sua vez, gastam anualmente cerca de mil milhões de euros na manutenção da sua política de multilinguismo, ou seja, na tradução de textos e na interpretação de comunicações orais.

O multilinguismo constitui sem dúvida um dos mais preciosos patrimónios da humanidade. Um mundo digital em que um único idioma viesse a assumir uma posição dominante, e viesse a substituir todos os outros, implicaria perdermos essa imensa riqueza imaterial que faz do mundo, em geral, e da Europa, em particular, um espaço único de encontro de culturas e diferenças.

É porém um fato, que não há vantagem em ignorar, que a diversidade linguística dificulta a comunicação do dia a dia. Apresenta-se como um obstáculo intransponível para os cidadãos, dificulta o debate político e atrasa o progresso económico e científico.

A tecnologia da linguagem e a investigação científica sobre as línguas naturais podem dar um contributo decisivo para se ultrapassarem estas barreiras linguísticas. No futuro, quando combinada com dispositivos e aplicações inteligentes, a tecnologia da linguagem ajudará falantes de diferentes línguas a comunicar naturalmente entre si. Preservando o multilinguismo, permitirá derrubar as fronteiras linguísticas que bloqueiam o acesso ao conhecimento, ajudando assim a concretizar todo o potencial da sociedade da informação.

Para atingir este objetivo, e preservar a diversidade cultural e linguística da Europa e do mundo, é necessário, antes de mais, fazer uma análise sistemática das particularidades linguísticas das diferentes línguas e do estado atual das tecnologias de apoio criadas para as mesmas. Essa é a finalidade do presente livro, no que diz respeito à língua portuguesa.

As ferramentas e aplicações para a tecnologia da linguagem e o processamento da fala atualmente existentes no mercado – dos sistemas de resposta a perguntas às interfaces em linguagem natural, incluindo as gramáticas computacionais ou as ferramentas de sumarização, entre muitas outras –, ainda estão porém muito distantes deste objetivo ambicioso. Isto aplica-se com particular acuidade à tradução automática, uma tecnologia especialmente relevante para a sustentabilidade do multilinguismo na era digital. Desde o final dos anos 70 que a União Europeia percebeu a extrema importância da tecnologia da linguagem como forma de contribuir para a unidade europeia e começou a financiar os primeiros projetos de investigação, como foi o caso do programa de tradução automática EUROTRA. Pela mesma altura, foram lançados projetos nacionais que produziram resultados assinaláveis mas que não conduziram a uma ação europeia concertada. Em contraste com este esforço de financiamento altamente seletivo, outras sociedades multilingues, como a Índia (22 línguas oficiais) ou a África do Sul (11 línguas oficiais), criaram recentemente programas nacionais de longo prazo para a investigação sobre a linguagem humana e o respetivo desenvolvimento tecnológico.

Nesta área, os atores dominantes são sobretudo empresas privadas, com fins lucrativos, sediadas na América do Norte. Estas empresas recorrem a abordagens estatísticas imprecisas que não utilizam métodos e conhecimentos linguísticos mais profundos. Por exemplo, as frases são automaticamente traduzidas através da comparação de uma nova frase com milhares de frases anteriormente traduzidas por seres humanos. Assim, a qualidade do resultado depende em grande medida da quantidade e da qualidade do corpus que serve de amostra. Embora a tradução automática de frases simples em línguas com uma quantidade suficiente de textos disponíveis possa alcançar resultados úteis, estes métodos estatísticos superficiais estão condenados ao fracasso no caso das línguas com um conjunto de material de amostra muito menor ou, sobretudo, no caso de frases com estruturas um pouco mais complexas.

Este livro fornece uma análise pormenorizada desta e de outras aplicações e soluções potenciadas pela tecnologia da linguagem. Como seria de esperar, e é revelado de forma circunstanciada nos volumes desta coleção de Livros Brancos, há diferenças dramáticas entre os vários países e as suas línguas no que diz respeito às soluções disponíveis e ao estado da investigação na área da ciência e tecnologia da linguagem.

O português é a quinta língua com maior número de falantes no mundo, com cerca de 220 milhões de falantes em quatro continentes – África, América, Ásia e Europa. Das línguas europeias, é a terceira língua com maior número de falantes no mundo. Face aos desafios colocados pela sociedade da informação num mundo globalizado, verifica-se a necessidade premente de se concentrarem mais esforços quer na criação de recursos linguísticos quer na investigação e desenvolvimento de ferramentas e aplicações para o processamento computacional do português.

O presente volume oferece uma exposição pormenorizada dos desafios, oportunidades e necessidades para o português na era digital. Uma das principais conclusões que resulta da análise feita neste livro é a de que o desenvolvimento de tecnologia da linguagem para a língua portuguesa é urgente e de importância fundamental para a consolidação do português como uma língua de comunicação internacional com projeção global.