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)

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.

2

LÍNGUAS EM RISCO: UM DESAFIO PARA A TECNOLOGIA DA LINGUAGEM

Somos testemunhas de uma revolução digital que está a ter um impacto radical na forma de comunicarmos e na sociedade em que vivemos. Os recentes desenvolvimentos nas áreas das Tecnologias da Informação e da Comunicação são por vezes comparados com a invenção da imprensa por Gutenberg.

O que pode esta analogia dizer-nos sobre o futuro da sociedade de informação europeia e sobre as nossas línguas em particular?

Na sequência da invenção da imprensa por Gutenberg, os avanços na comunicação e na partilha de conhecimentos foram concretizados através de inúmeras realizações, das quais a tradução da Bíblia do Latim para as línguas vernáculas da Europa é apenas um dos aspetos mais reconhecidos. Nos séculos seguintes, foram desenvolvidas novas técnicas para melhor lidar com o processamento da linguagem e a partilha de conhecimento:

- a padronização ortográfica e gramatical das principais línguas permitiu a rápida divulgação de novas perspetivas científicas e intelectuais;
- o desenvolvimento das línguas oficiais tornou possível aos cidadãos comunicarem dentro de certas fronteiras (muitas vezes políticas);
- o ensino e a tradução de línguas permitiram uma partilha de conhecimento entre línguas;
- a criação de diretrizes editoriais e bibliográficas garantiu a qualidade e a disponibilidade do material impresso;
- o surgimento de diferentes meios de comunicação, como jornais, rádio, televisão, livros e outros suportes e formatos, veio dar resposta às diferentes necessidades de comunicação.

Estamos a testemunhar uma revolução digital com um impacto que tem sido comparado com invenção da imprensa por Gutenberg.

De forma análoga, nos últimos vinte anos, as Tecnologias da Informação e da Comunicação vieram ajudar ainda mais a automatizar e a facilitar o processamento da linguagem e a comunicação:

- as aplicações para edição de texto (*desktop publishing software*) substituem a datilografia e a composição tipográfica;
- as projeções de transparências são substituídas por apresentações em Powerpoint;
- o correio eletrónico permite receber e enviar documentos de forma mais rápida que o fax;
- o Skype permite realizar chamadas de telefone gratuitas ou a preços reduzidos pela internet, assim como videoconferências;
- os formatos de codificação de áudio e vídeo facilitam a troca de conteúdos multimédia;
- os motores de busca permitem aceder a informação com base em palavras-chave;

- os serviços de tradução online, como o Google Translate, produzem traduções rápidas ainda que apenas aproximadas;
- as plataformas de redes sociais como o Facebook, o Twitter ou o Google+ facilitam a comunicação, a colaboração e a partilha de informação.

Apesar de estas ferramentas e aplicações serem úteis, ainda não são capazes de apoiar, de forma sustentada, uma sociedade europeia multilingue para todos, onde a informação e os bens possam circular livremente.

2.1 FRONTEIRAS LINGUÍSTICAS ENTRAVAM A SOCIEDADE DE INFORMAÇÃO EUROPEIA

Não podemos saber exatamente como será o futuro da sociedade de informação. Há porém uma forte probabilidade de que a revolução nas tecnologias da comunicação venha a aproximar, de forma inovadora, pessoas que falam diferentes línguas. Esta situação vai pressionar toda a gente a aprender novas línguas e pressiona sobretudo os criadores de software a desenvolverem novas aplicações que permitam a inter-compreensão entre falantes de diferentes idiomas e o acesso a conhecimento partilhado. Este espaço económico e de informação global envolve a interação entre línguas, falantes e conteúdos no âmbito de novos meios de comunicação. A recente popularidade das redes sociais (Wikipédia, Facebook, Twitter, YouTube e, mais recentemente, o Google+) é apenas a ponta visível de um iceberg.

A economia e o espaço de informação globais colocam-nos perante mais línguas, falantes e conteúdos.

Hoje, podemos transmitir gigabytes de texto para todo o mundo em poucos segundos antes ainda de nos conseguirmos aperceber de que o conteúdo está redigido numa língua que não entendemos. De acordo com um recente relatório da Comissão Europeia, 57% dos utilizadores da internet compram bens e serviços em línguas que não a sua (o inglês é a língua estrangeira mais usada, seguido pelo francês, alemão e espanhol). Por sua vez, 55% dos utilizadores leem conteúdos numa língua estrangeira, enquanto apenas 35% utilizam outra língua para escrever mensagens de correio eletrónico ou colocar comentários na internet [2].

Há alguns anos atrás, o inglês era a língua franca na internet – a maior parte dos conteúdos estavam de facto em inglês – mas agora a situação mudou radicalmente. A quantidade de conteúdos online noutras línguas europeias (assim como em línguas asiáticas e do Próximo Oriente) aumentou exponencialmente.

Surpreendentemente, esta divisão digital criada pelas fronteiras linguísticas não recebe muita atenção pública. Ainda assim, levanta uma questão premente:

Que línguas europeias vão prosperar na informação em rede e na sociedade do conhecimento, e quais estão condenadas a desaparecer?

2.2 AS NOSSAS LÍNGUAS EM RISCO

Embora a imprensa escrita tenha ajudado a intensificar a troca de informação na Europa, também levou à extinção de muitas línguas europeias. Línguas regionais e minoritárias raramente foram impressas, como o Cornish e o Dálmata, e foram reduzidas a formas orais de transmissão, o que limitou o seu uso.

No futuro, terá a internet o mesmo impacto nas nossas línguas?

As cerca de 80 línguas da Europa são um dos mais ricos e importantes patrimónios culturais e uma parte vital do seu modelo social, que é único [3]. Enquanto línguas como o inglês e o espanhol sobreviverão no mercado

digital emergente, muitas línguas europeias poderão tornar-se irrelevantes numa sociedade ligada em rede. Isso enfraqueceria a posição global da Europa e iria contra o objetivo estratégico da participação de todos os cidadãos europeus em igualdade de circunstâncias, independentemente da sua língua.

A grande variedade de línguas na Europa é um dos seus patrimónios culturais mais ricos e importantes.

De acordo com um relatório da UNESCO sobre multilinguismo, as línguas são um meio essencial para o exercício dos direitos fundamentais, como a expressão política, a educação e a participação social [4].

2.3 A TECNOLOGIA DA LINGUAGEM É UMA TECNOLOGIA FACILITADORA

No passado, os esforços de investimento para a preservação das línguas concentraram-se no ensino e na tradução. De acordo com uma estimativa, o mercado europeu de tradução, interpretação, localização de software e preparação de websites para o mercado global foi de 8,4 mil milhões de euros em 2008 e deverá crescer 10% por ano [5]. No entanto, este número abrange apenas uma pequena parte das necessidades atuais e futuras da comunicação entre línguas.

A solução mais viável para garantir uma utilização ampla e continuada das várias línguas na Europa do futuro encontra-se no recurso a tecnologia apropriada, tal como recorremos a tecnologia apropriada para dar resposta às nossas necessidades, por exemplo, nas áreas da energia e dos transportes, ou para apoiar cidadãos com necessidades especiais, entre tantos outros casos.

A tecnologia da linguagem, dirigida a todas as formas de texto escrito e discurso falado, ajuda as pessoas a colaborar, a concretizar negócios, a partilhar conhecimentos e a participar em debates sociais e políticos, independentemente das barreiras linguísticas e das aptidões informáticas de cada um.

A tecnologia da linguagem funciona muitas vezes "nos bastidores", de forma invisível dentro de sistemas de software complexos, ajudando-nos já hoje em dia em tarefas como:

- encontrar informação com um motor de busca;
- verificar a ortografia e a gramática com um processador de texto;
- ver as recomendações para um produto numa loja online;
- seguir as indicações verbais de um sistema de navegação;
- traduzir páginas web com um serviço online.

A tecnologia da linguagem consiste num conjunto de aplicações nucleares que permitem uma série de procedimentos embebidos em sistemas mais amplos. Um dos objetivos desta coleção de Livros Brancos da META-NET é o de perceber o nível de desenvolvimento desta tecnologia para cada uma das línguas europeias.

A Europa precisa de tecnologia da linguagem robusta e económica para todas as línguas europeias.

Para manter a sua posição na linha da frente da inovação mundial, a Europa necessitará de tecnologia da linguagem que esteja adaptada a todas as línguas europeias e que seja igualmente robusta e económica, e bem integrada em ambientes de software-chave.

Sem tecnologia da linguagem suficientemente desenvolvida, não nos será possível alcançar uma experiência efetivamente interativa, multimédia e multilingue num futuro próximo.

2.4 OPORTUNIDADES PARA A TECNOLOGIA DA LINGUAGEM

O desenvolvimento da imprensa, com a duplicação rápida de uma imagem de texto, constituiu um avanço tecnológico fundamental. Mas os seres humanos continuam ainda a ter de fazer o trabalho árduo de buscar, apreciar, traduzir e resumir a informação.

A tecnologia da linguagem pode agora simplificar e automatizar muitos dos processos de tradução, produção de conteúdos e gestão de conhecimentos. Permite igualmente desenvolver interfaces de voz para eletrodomésticos, máquinas, veículos, computadores e robôs. As aplicações industriais e comerciais ainda estão num estádio inicial de desenvolvimento, mas os resultados em Investigação e Desenvolvimento estão a criar uma janela de oportunidade genuína. Por exemplo, a tradução automática já é razoavelmente precisa em certos domínios específicos e algumas aplicações experimentais já asseguram informação multilingue e gestão do conhecimento, assim como a possibilidade de produzir conteúdos, em várias línguas europeias.

Tal como a maioria das tecnologias, as primeiras aplicações para a linguagem humana, como as interfaces com o utilizador baseadas na voz ou os sistemas de diálogo, foram desenvolvidas para domínios altamente especializados, e em regra apresentam limitações de desempenho. Contudo, existem imensas oportunidades de mercado nas indústrias da educação e do entretenimento para a integração da tecnologia da linguagem em jogos, pacotes de jogos educativos, bibliotecas, ambientes de simulação ou programas de formação. Os serviços de informação móveis, os programas de aprendizagem de uma língua assistida por computador, os ambientes de e-learning, as ferramentas de autoavaliação e os programas de deteção de plágio são apenas alguns dos exemplos onde esta tecnologia pode desempenhar um papel importante. A popularidade das redes sociais, como o Twitter e o Facebook, sugerem uma maior necessidade de sofisticação da tecnologia da linguagem para se poder monitorizar mensagens, resumir discussões, sugerir tendências de opinião, detetar respostas emocionais, identificar infrações aos direitos de autor ou encontrar usos indevidos.

> A tecnologia da linguagem ajuda a superar os obstáculos colocados pela diversidade linguística.

A tecnologia da linguagem representa uma enorme oportunidade para a União Europeia. Pode ajudar a resolver a complexa questão do multilinguismo na Europa, nomeadamente ajudando a que diferentes línguas coexistam naturalmente nos negócios, nas organizações e nas escolas. Os cidadãos têm a necessidade de comunicar para além destas fronteiras linguísticas que cruzam o Mercado Comum Europeu e a tecnologia da linguagem pode assim ajudar a superar os obstáculos que ainda existem, permitindo o uso livre e ilimitado do idioma de cada um.

Pensando a longo prazo, a tecnologia da linguagem multilingue europeia poderá ser inclusive uma referência inovadora para os nossos parceiros globais e as suas comunidades multilingues.

A tecnologia da linguagem pode ser vista como uma forma de "tecnologia de apoio" que ajuda a ultrapassar os obstáculos da diversidade linguística e tornar as comunidades linguísticas mais acessíveis umas às outras.

2.5 DESAFIOS PARA A TECNOLOGIA DA LINGUAGEM

Apesar do progresso assinalável na área da tecnologia da linguagem nos últimos anos, o atual ritmo de progresso tecnológico e de inovação em termos de produtos é demasiado lento. As tecnologias com maior utilização,

como os corretores ortográficos e gramaticais em processadores de texto, são normalmente monolingues e estão disponíveis apenas para um pequeno número de idiomas. Os serviços de tradução automática online, apesar de serem úteis para gerar rapidamente uma aproximação razoável ao conteúdo de um documento, veem-se enredados em imensa dificuldade quando lhe são pedidas traduções mais precisas e completas.

O ritmo atual do progresso da tecnologia da linguagem é demasiado lento.

Devido à complexidade da linguagem humana, providenciar a modelação computacional dos nossos idiomas e testá-la no mundo real é um processo longo e oneroso, que exige compromissos de financiamento sustentados. A Europa tem, por isso, de manter o seu papel pioneiro de lidar com os desafios tecnológicos colocados por uma comunidade multilingue, inventando novos métodos para acelerar o desenvolvimento de forma pervasiva.

2.6 AQUISIÇÃO DA LINGUAGEM POR SERES HUMANOS E POR MÁQUINAS

Para ilustrar como os computadores lidam com a linguagem natural e as razões pelas quais é difícil programá-los para esse efeito, vamo-nos centrar, muito brevemente, na forma como os seres humanos adquirem as suas primeira e segunda línguas, e depois ver como funcionam os sistemas de tecnologia da linguagem.

Os seres humanos adquirem competências linguísticas de dois modos diferentes. Os bebés aprendem uma língua interagindo linguisticamente e ouvindo as interações entre os pais, irmãos e outros membros da família. Por volta dos dois anos de idade, as crianças começam a produzir as suas primeiras palavras e frases curtas. Isto só é possível porque os seres humanos têm uma predisposição genética para imitar e racionalizar o que ouvem. Aprender uma segunda língua numa idade mais avançada exige um maior esforço cognitivo, sobretudo quando quem aprende não está inserido numa comunidade de falantes dessa língua. Na escola, as línguas estrangeiras são normalmente adquiridas através do ensino da estrutura gramatical, vocabulário e ortografia, utilizando exercícios que descrevem conhecimentos linguísticos em termos de regras abstratas, tabelas e exemplos.

Os seres humanos adquirem aptidões linguísticas de dois modos diferentes: aprendendo a partir de exemplos e aprendendo as regras subjacentes.

Passando agora para a tecnologia da linguagem, os dois tipos principais de sistemas adquirem capacidades linguísticas de forma similar. As abordagens estatísticas permitem obter conhecimentos linguísticos a partir de vastas coleções de exemplos concretos de textos. Embora seja suficiente usar textos numa única língua para, por exemplo, treinar um corretor ortográfico, são necessários textos paralelos em duas ou mais línguas para o treino de um sistema de tradução automática. O algoritmo de aprendizagem automática pode então adquirir os padrões quanto ao modo como as palavras, expressões e frases completas são traduzidas.

Em regra, esta abordagem estatística requer milhões de frases para se obter um acréscimo significativo da qualidade no seu desempenho. Esta é uma das razões por que os fornecedores de motores de busca pretendem recolher o máximo de material escrito possível. Por exemplo, a correção ortográfica em processadores de texto ou serviços como o Google Search ou o Google Translate depende de abordagens estatísticas. A grande vantagem da estatística é que a máquina realiza uma rápida aprendizagem em séries contínuas de ciclos de treino.

Uma outra abordagem na tecnologia da linguagem, em geral, e na tradução automática, em particular, consiste na construção de sistemas baseados em regras. Peritos nas áreas da Linguística, Linguística Computacional e Engenharia Informática têm de, primeiro, codificar a análise gramatical (regras gramaticais) e compilar listas de vocabulário (léxicos). Isto requer imenso tempo e trabalho. Alguns dos principais sistemas de tradução automática baseados em regras têm estado em constante desenvolvimento desde há mais de 20 anos. A grande vantagem de sistemas baseados em regras é que os peritos têm um controlo mais pormenorizado sobre o processamento da linguagem. Isto torna possível corrigir de forma sistemática os erros no software e dar uma resposta detalhada ao utilizador, especialmente quando os sistemas baseados em regras são usados para a aprendizagem de línguas. Contudo, devido ao alto custo deste trabalho, a tecnologia da linguagem baseada em regras tem sido desenvolvida apenas para alguns idiomas até agora.

Como os pontos fortes e fracos de sistemas baseados em estatística e em regras tendem a ser complementares, a investigação atual concentra-se em abordagens híbridas que combinem as duas metodologias. No entanto, até agora, estas abordagens têm tido menos sucesso nas aplicações industriais do que nos laboratórios de investigação.

Os dois principais tipos de tecnologia da linguagem adquirem capacidades de processamento de uma forma algo similar à forma como os seres humanos o fazem.

Como vimos neste capítulo, muitas aplicações amplamente utilizadas na atual sociedade de informação dependem fortemente da tecnologia da linguagem. Devido à sua comunidade multilingue, isto é particularmente verdadeiro no espaço económico e de informação da Europa. Embora a tecnologia da linguagem tenha obtido progressos assinaláveis nos últimos anos, há ainda um enorme potencial para melhorar os resultados alcançados. Nos próximos capítulos, vamos descrever o papel do português na sociedade europeia de informação e no mundo e avaliar o estado atual da tecnologia da linguagem para a língua portuguesa.

3

O PORTUGUÊS NA SOCIEDADE DE INFORMAÇÃO

3.1 FACTOS GERAIS

O português é a terceira língua europeia com maior número de falantes no mundo, com cerca de 220 milhões de falantes em quatro continentes, dos quais 200 milhões têm o português como língua materna: África, América, Ásia e Europa [6, 7]. É a língua oficial de Angola, Brasil, Cabo Verde, Guiné-Bissau, Macau, Moçambique, Portugal, São Tomé e Príncipe, Timor-Leste, e desde 2010, da Guiné Equatorial.

O português é a terceira língua europeia mais falada no mundo, com cerca de 220 milhões de falantes.

Em resultado de movimentos migratórios [8, 9], o português é também falado por comunidades presentes em muitos países, ocupando em alguns deles uma importante posição entre a população estrangeira. É o caso, na Europa, do Luxemburgo (cerca de 25% da população), Andorra (à volta de 11%), França, Alemanha, Reino Unido, Suíça, Espanha e Bélgica [10].

O português é uma das línguas oficiais da União Europeia, do Mercosul e da União Africana. Com o avanço da alfabetização nos países africanos e em Timor-Leste, o português tem um grande potencial de crescimento.

As expedições e o comércio costeiro que Portugal manteve durante vários séculos apresentam hoje contrapartidas linguísticas: o português incorporou palavras de origem africana, ameríndia e asiática, mas também deu a sua contribuição lexical para muitas línguas no mundo e vários pidgins e crioulos do Oceano Atlântico, Oceano Pacífico e Oceano Índico [11, 12].

Em Portugal, a divisão geográfica dos dialetos [13] distingue os dialetos do Centro-Sul, os dialetos do Norte e os dialetos das ilhas atlânticas. Os dialetos do Norte podem ser identificados pela ausência da distinção fonológica entre /b/ e /v/, com prevalência do /b/, pela preservação de antigos ditongos, e pela existência de fricativas ápicoalveolares. As diferenças entre estes dialetos encontram-se sobretudo ao nível da fonética e fonologia e ao nível lexical, sendo todos eles mutuamente compreensíveis de forma imediata (possivelmente com a exceção de alguns dialetos das ilhas).

Quanto ao Brasil, dada a dimensão geográfica deste país, não é viável apresentar aqui as suas variedades linguísticas. Por razões geográficas, políticas e sociais, não é possível falar de uma variedade padrão do português do Brasil. Os especialistas tendem a mencionar "normas urbanas cultas".

A situação das variedades africanas do português é variada: enquanto em Angola e Moçambique o número de falantes de português tem vindo a aumentar desde a independência destes países, noutros casos, como São Tomé e Princípe ou Cabo Verde, em muitas circunstâncias utiliza-se amplamente o crioulo e o português é adquirido como língua segunda.

Todas as variantes do português nos diferentes continentes são mutuamente compreensíveis de forma generalizada.

3.2 PARTICULARIDADES DA LÍNGUA PORTUGUESA

O português é uma língua românica [14], pelo que a maioria do seu léxico deriva do Latim. Em diferentes momentos da sua história, integrou muitas palavras de várias outras línguas, as quais, em muitos casos, permanecem entre os vocábulos mais frequentes. Exemplos pré-latinos: *barranco*, *seara*, *bruxa*; germânicos: *luvas*, *bando*, *guerra*; árabes: *aldeia*, *açúcar*, *laranja*; africanos: *batuque*, *inhame*; asiáticos: *chá*, *biombo*, *bengala*; e ameríndios: *cacau*, *tapioca*. As línguas dos povos com os quais os portugueses estabeleceram contactos durante a expansão marítima também integraram palavras portuguesas, como, no caso do japonês, as palavras *bidoro* (do português *vidro*) e *pan* (do português *pão*).

Para um ouvinte que não domina a língua portuguesa, a variante europeia desta língua pode muitas vezes soar como uma sequência de consoantes. Isto deve-se ao facto de as vogais átonas do português serem muitas vezes enfraquecidas ou mesmo não realizadas, ao invés do que acontece com outras línguas românicas. Este processo fonológico do enfraquecimento das vogais é uma mudança tardia no português europeu e não teve lugar na variedade falada no Brasil, a qual, deste ponto de vista, se encontrará mais próxima do português falado há séculos atrás.

O português é uma lingua românica. Ao longo da sua história, integrou muitas palavras de outras línguas.

A ordem básica das palavras em português é dita ser SVO – Sujeito Verbo Objeto (*ele leu o livro*). Em alguns contextos pragmáticos, como por exemplo contextos enfáticos, a ordem VSO pode ocorrer (*lês tu o livro*) e as ordens OSV ou OVS são possíveis em construções que na terminologia gramatical são ditas marcadas (*o livro, ele não leu*).

O português é uma língua que permite sujeitos nulos, isto é, o sujeito de uma dada frase pode não estar realizado foneticamente (_ *li o livro*). Quando o sujeito tem a flexão de primeira pessoa, a sua não realização fonética é a opção por omissão. Adicionalmente, em regra, não ocorrem pronomes expletivos nas construções impessoais (_ *há um livro sobre esse tema*). Esta é uma das características do português que representa um desafio acrescido para a análise sintática automática dos textos e da fala.

O paradigma flexional do português é muito mais rico que o de línguas como o inglês, em particular no que diz respeito aos verbos. Por exemplo, um verbo pode ter diferentes marcas para aspeto, tempo, modo, pessoa, número, género ou polaridade, atingindo mais de 160 formas flexionadas diferentes, incluindo as simples e compostas [15].

Algumas propriedades da língua portuguesa constituem um desafio acrescido para a tecnologia da linguagem.

Além disso, há dois paradigmas de flexão verbal que não existem em outras línguas românicas e que são muito frequentes em português: o infinitivo flexionado e o futuro do conjuntivo. O primeiro partilha o tema com o infinitivo não flexionado (por exemplo, *cantar*) ao qual se juntam marcadores flexionais de aspeto, tempo, modo, pessoa e número (por exemplo, *para tu cantares*). Exceto no caso dos verbos irregulares, as formas flexionadas do futuro do conjuntivo são homónimas com as do infinitivo não flexionado, o que aumenta o número de formas ambíguas no paradigma flexional do verbo.

A posição dos pronomes clíticos na frase é outra característica que coloca desafios específicos ao processamento automático da língua portuguesa. Os pronomes clíticos podem ocorrer antes ou depois do verbo, exceto nos tempos futuro e condicional, em que podem ocorrer antes ou no meio da forma verbal (*dar-lho-ei*). A presença de um clítico de terceira pessoa no meio ou após o verbo pode afetar a forma do próprio verbo. Por exemplo, na sequência final *-ar*, o *-r* cai e a vogal é acentuada (*dá-lo-ei*).

3.3 DESENVOLVIMENTOS RECENTES

Sendo o inglês a língua mais difundida no mundo, a sua influência noutras línguas, incluindo o português, é cada vez mais notória. O cinema e a televisão, sobretudo séries norte-americanas, a música e a internet, contribuem para a presença regular da língua inglesa no quotidiano e muitas palavras desta língua acabam por ser integradas no português.

É sobretudo em línguas de especialidade, como a gestão ou a informática, que as palavras inglesas ganham maior visibilidade, como *CEO*, *manager*, *briefing*, *casual day* ou *download*, *pen USB*, *upload*, *online* ou *site*, e também *lifting*, *e-learning* ou *shopping*, entre muitas outras.

No que diz respeito à música, embora haja muitos projetos musicais com letras em inglês dirigidos a um público mais jovem, a música cantada em português, incluindo o fado e outros tipos de música tradicional portuguesa, está agora a recuperar uma grande audiência de todas as idades.

Na última década, tem havido um crescimento da relevância do português no contexto económico internacional, sobretudo devido ao desenvolvimento económico do Brasil e dos países africanos de língua oficial portuguesa. No âmbito das Nações Unidas, o português tem desempenhado um papel cada vez mais importante, com iniciativas para torná-lo uma das línguas de trabalho, como já acontece na União Europeia e no Mercosul.

A crescente importância do português a nível internacional reflete-se no número crescente de pessoas que se inscrevem em cursos de português por todo o mundo.

3.4 DIVULGAÇÃO E PROMOÇÃO

A Comunidade dos Países de Língua Oficial Portuguesa (CPLP) é uma organização intergovernamental para a cooperação. Um dos seus objetivos consiste na divulgação e promoção do português. O Instituto Internacional da Língua Portuguesa é o organismo da CPLP especificamente dedicado à promoção da língua portuguesa como língua internacional de projeção global. Foi também no seio da CPLP que foram empreendidos esforços conducentes ao Novo Acordo Ortográfico [16], comum a todos os países desta comunidade, de forma a apoiar a consolidação da língua no cenário económico e político internacional. Este Novo Acordo Ortográfico abrange todos os países de língua oficial portuguesa.

A Comunidade dos Países de Língua Oficial Portuguesa (CPLP) é uma organização intergovernamental com um papel ativo na divulgação e promoção da Língua Portuguesa.

A Academia das Ciências de Lisboa e a Academia Brasileira das Letras dedicam-se à divulgação da língua portuguesa, nomeadamente através da edição de dicionários de referência: o Dicionário da Língua Portuguesa Contemporânea, no caso da Academia portuguesa, e o Dicionário da Academia Brasileira de Letras, no caso da Academia brasileira.

O Instituto Camões é uma instituição sob a tutela do Ministério dos Negócios Estrangeiros de Portugal e um

dos seus principais objetivos é a promoção do português no mundo. Esta instituição coordena e apoia o ensino do português em universidades e centros de cultura e língua portuguesa em todo o mundo. Concede financiamento a atividades culturais relacionadas com a língua, concedendo bolsas de estudo a nacionais e estrangeiros e apoiando o português como língua de comunicação internacional, particularmente em instituições internacionais como as Nações Unidas.

O Instituto Camões é a instituição sob a tutela do Ministério dos Negócios Estrangeiros de Portugal que tem por missão promover a língua portuguesa.

A Fundação Calouste Gulbenkian [17], sediada em Lisboa, também apoia a promoção da língua portuguesa. Por exemplo, através do seu serviço internacional, equipa Departamentos de Português e História em universidades estrangeiras ou instituições culturais de todo o mundo com livros de autores portugueses. Financia a organização de congressos, conferências e seminários sobre língua e literatura portuguesas. Financia também projetos de investigação, como por exemplo, o projeto do Corpus de Referência do Português Contemporâneo ou o projeto Gramática do Português do Centro de Linguística da Universidade de Lisboa.

Nos últimos anos, o Brasil tem aumentado a cooperação internacional, com especial incidência no domínio da educação, com reflexos no apoio à língua portuguesa. Neste sentido, existem acordos com Angola e Moçambique para a oferta de cursos de pós-graduação in loco e à distância. Já com países de língua espanhola que fazem fronteira com o Brasil, como o Uruguai, existem bolsas de estudo para docentes das principais universidades e, nas zonas fronteiriças desses mesmos países, está a ser estimulada a educação bilingue.

A rádio e televisão públicas de Portugal têm-se empenhado na promoção do português através da transmissão de programas de divulgação que procuram ensinar boas práticas no uso da língua portuguesa, emitindo diariamente programas para esclarecer algumas dúvidas frequentes sobre a norma do português. Na cadeia de televisão pública, o programa semanal Cuidado com a Língua é simultaneamente educativo e divertido e ajuda a divulgar o Novo Acordo Ortográfico. Na rádio pública, há debates regulares sobre as boas práticas do português escrito e falado. Tem havido também muitas publicações dedicadas à língua portuguesa, procurando atrair mais público para o seu uso adequado. Todos estes programas e publicações visam responder a um interesse empenhado da população pelas questões da língua. Também as estações de rádio e televisão em língua portuguesa, dispersas pelo mundo, têm feito um trabalho assinalável para manter o uso do português junto dos emigrantes e dos seus descendentes.

O novo Acordo Ortográfico para o português foi aprovado no quadro da Comunidade dos Países de Língua Oficial Portuguesa (CPLP).

No setor da música, o uso do português tem sido apoiado através de um sistema de quotas nas rádios portuguesas. A lei estipula uma percentagem obrigatória, nomeadamente 25%, de música em português nos programas emitidos.

A língua portuguesa também é promovida através do aumento da projeção internacional de autores africanos, brasileiros e portugueses. Pode-se destacar filósofos portugueses, como Eduardo Lourenço ou Fernando Gil, assim como escritores portugueses, como António Lobo Antunes ou José Saramago, o recentemente desaparecido Prémio Nobel da Literatura, cujas obras se encontram traduzidas em todo o mundo, entre vários outros. No contexto da literatura brasileira, Jorge Amado

ou Paulo Coelho são exemplos de escritores com ampla tradução e divulgação a nível mundial. No que diz respeito aos autores africanos, Mia Couto, de Moçambique, e José Eduardo Agualusa ou Luandino Vieira, de Angola, são alguns exemplos também a merecer destaque.

3.5 LÍNGUA PORTUGUESA E EDUCAÇÃO

Nos últimos anos, teve lugar em Portugal um grande investimento no desenvolvimento de uma rede de bibliotecas escolares. Isto foi feito no âmbito do Plano Nacional de Leitura, cujo objetivo é a melhoria dos índices de literacia dos estudantes portugueses nos vários níveis de aprendizagem, com especial enfoque nos primeiros anos de ensino. Também no Brasil têm sido implementadas, de forma gradual, políticas educativas que permitam um aumento do nível de literacia entre os estudantes brasileiros.

Outra iniciativa recente em Portugal foi a integração generalizada das novas tecnologias da informação nas escolas. Os alunos mais novos têm a possibilidade de adquirir a baixo custo, e nalguns casos gratuitamente, computadores portáteis especialmente concebidos para os diferentes níveis de ensino. Em conjunto com este acesso a computadores pessoais, foram desenvolvidos programas educativos em português que estimulam, entre outros aspetos, a aprendizagem da gramática.

O Plano Nacional de Leitura em Portugal tem como objetivo a promoção dos índices de literacia dos estudantes. Iniciativas semelhantes têm sido desenvolvidas no Brasil.

Cabe também referir o papel desempenhado pela Fundação Gulbenkian, nomeadamente no apoio dado à constituição de bibliotecas escolares e públicas. Relevante foi também o apoio dado ao projeto Diversidade Linguística na Escola Portuguesa [18], desenvolvido em conjunto com o Instituto de Linguística Teórica e Computacional, e cujo principal objetivo é o de contribuir para a integração escolar de alunos que não têm o português como língua materna.

Os recentes resultados do PISA 2009 (Programme for International Student Assessment) demonstraram uma melhoria comparativa dos resultados dos alunos portugueses ao nível da leitura, das ciências e da matemática, com especial destaque para a componente da leitura.

Num futuro próximo, espera-se o continuado impacto benéfico deste investimento no Plano Nacional de Leitura e nas novas tecnologias, assim como da recente medida de aumentar a escolaridade obrigatória para doze anos.

3.6 ASPETOS INTERNACIONAIS

Na sequência das explorações marítimas portuguesas, das descobertas geográficas e da abertura de novas rotas no comércio mundial, desde há séculos que a língua portuguesa tem sido projetada em todo o mundo como uma das línguas mais importantes para o comércio e para os negócios.

O português é atualmente uma língua de comunicação internacional com projeção global, com cerca de 220 milhões de falantes, dos quais cerca de doze milhões encontram-se na Europa, com cerca de 10 milhões em Portugal [19]. É no Brasil que se encontra a maior comunidade de falantes do português, com 190 milhões. A par do tamanho da sua população, o Brasil está a contribuir para uma cada vez maior projeção internacional da língua portuguesa em resultado do seu desenvolvimento económico e da sua posição na cena internacional como uma das potências emergentes do século XXI. Tem-se registado um interesse crescente pela língua portuguesa, sendo o português ensinado em muitos países do mundo [20]. Diversas Câmaras de Comércio têm proporcionado aulas de português para potenciais

investidores, como foi o caso recente da Câmara Italiana em Portugal, só para citar um exemplo entre muitos outros. As comunidades de emigrantes portugueses no mundo têm sido outro fator de promoção do ensino do português.

Há um crescente interesse pela língua portuguesa no mundo, tanto no setor académico como no setor da economia e dos negócios.

A língua portuguesa é atualmente uma das 23 línguas oficiais da União Europeia e tem sido incluída em muitos projetos de investigação financiados pela Comissão Europeia com o objetivo de se desenvolver recursos e tecnologia da linguagem. A língua portuguesa é também língua administrativa e de trabalho de 27 organizações internacionais, incluindo, por exemplo, a Comunidade dos Países de Língua Oficial Portuguesa (CPLP), o Mercosul, a União Latina ou a Federação Internacional de Futebol (FIFA).

A língua portuguesa é língua administrativa e de trabalho de 27 organizações internacionais.

Apesar da sua progressiva projeção, a língua portuguesa pode enfrentar alguns desafios no que toca à sua posição como língua de comunicação internacional. Na América Latina, com cerca de 190 milhões de falantes, o português co-existe com grandes comunidades de falantes de espanhol. Na Europa, um continente multilingue, o português conta apenas com cerca de doze milhões de falantes, incluindo as comunidades emigrantes. Na Ásia, é língua oficial somente em Timor-Leste e Macau. E em África, a par do facto de muitas línguas nativas co-existirem com o português, o inglês e o francês são línguas com uma projeção forte e concorrente nesse continente.

3.7 A LÍNGUA PORTUGUESA NA INTERNET

Um apanhado geral dos dados estatísticos sobre a língua portuguesa revela que esta é uma das línguas mais utilizadas na internet. De acordo com estimativas recentes, o português é a quinta língua mais usada na internet, sendo ultraprassada apenas pelo inglês, chinês, espanhol e japonês [21]. Esta pesquisa mostra que cerca de 82,5 milhões de utilizadores usam o português para navegar na internet, e que numa década, entre 2000 e 2010, o número de utilizadores que usam o português registou uma surpreendente expansão de 990%.

O português está particularmente bem posicionado quando se trata da presença nas redes sociais. Um estudo semântico e quantitativo de 2,8 milhões de tweets, realizado pela Semiocast, revela que o português é a terceira língua mais usada no Twitter, depois do inglês e do japonês [22].

A língua portuguesa é a quinta mais utilizada na internet, onde registou um surpreendente crescimento de 990% na última década.

Isto resulta do enorme aumento do acesso à internet no Brasil, particularmente entre os jovens. Este país tem um dos maiores números de utilizadores de internet em todo o mundo, com 72 milhões de internautas [23], e as respostas a um questionário do censos revelaram que o número de utilizadores da internet com 10 anos ou mais deu um salto de doze milhões desde 2008 [24]. Portugal, por sua vez, tem cerca de 5 milhões de utilizadores da internet [25, 26] e as estatísticas revelam que o número de subscritores de acesso à internet tem registado um aumento notório: em 2001 havia pouco menos de 500 mil assinantes, e as últimas estimativas indicam perto de 2 milhões de assinantes atualmente [27]. As estatísticas revelam também que em 2006, 95% das empresas

com dez ou mais funcionários usavam computador, enquanto 84% utilizavam o email e 83% tinham acesso à internet; que em 2008, mais de 90% dos indivíduos com idades entre os 10 e 15 anos usavam computador (96,6%) e a internet (92,7%); e que em 2010, 54% dos lares portugueses tinham acesso à internet [27].

Paralelamente ao esforço de assegurar a presença de institutos, agências e serviços públicos na internet, em 2007, foi implementado em Portugal o Plano Nacional para a Promoção da Acessibilidade [28], orientado para promover a inclusão social através da Sociedade de Informação e para permitir o acesso a conteúdos na rede por parte de cidadãos com deficiência.

É pois inequívoco o uso crescente da língua portuguesa na internet.

A par dos dados acima apresentados, vale a pena realçar que o português está presente em vários sites de instituições políticas e económicas internacionais, como os sites da União Europeia ou do Mercosul, só para citar dois exemplos, devendo ser dada continuidade aos esforços para que esta língua esteja presente noutras instituições onde ainda não é opção.

4

TECNOLOGIA DA LINGUAGEM PARA O PORTUGUÊS

A tecnologia da linguagem é usada para desenvolver sistemas de software cujo objetivo é lidar com a linguagem humana, pelo que frequentemente é também designada por tecnologia da linguagem humana.

A linguagem humana surge na forma falada e escrita. Enquanto a fala representa a forma de comunicação mais antiga em termos de evolução humana, e o meio de comunicação mais natural, é através dos textos escritos que se transmite informação complexa e é neles que está armazenada a maioria do conhecimento humano. As tecnologias de processamento da fala e do texto analizam ou produzem linguagem, sob estas diferentes formas, através da utilização de dicionários, regras de gramática e semânticas. Isto significa que a tecnologia da linguagem liga a linguagem a várias formas de conhecimento, independentemente do meio (textual ou oral) em que é expressa.

Quando comunicamos, combinamos a linguagem com outras formas de comunicação e outros meios de informação. Falar pode envolver gestos e expressões faciais. Os textos digitais são acompanhados por imagens e sons. Os filmes podem incluir linguagem sob a forma oral ou escrita. Isto quer dizer que as tecnologias da fala e do texto se entrecruzam com outras tecnologias de modo a facilitar o processamento da comunicação multimodal.

A Figura 1 apresenta, em traços muitos gerais, este enquadramento da tecnologia da linguagem.

Neste capítulo, começar-se-á por apresentar as áreas de aplicações nucleares para a tecnologia da linguagem, descrevendo sumariamente o seu estado de desenvolvimento. No final, apresentar-se-á uma apreciação no que respeita ao estado de desenvolvimento da tecnologia da linguagem para o português. Isto permitirá obter uma perspetiva sobre o estado da arte desta tecnologia para a língua portuguesa e uma comparação sinóptica com o que se passa relativamente às outras línguas abrangidas por esta coleção de Livros Brancos.

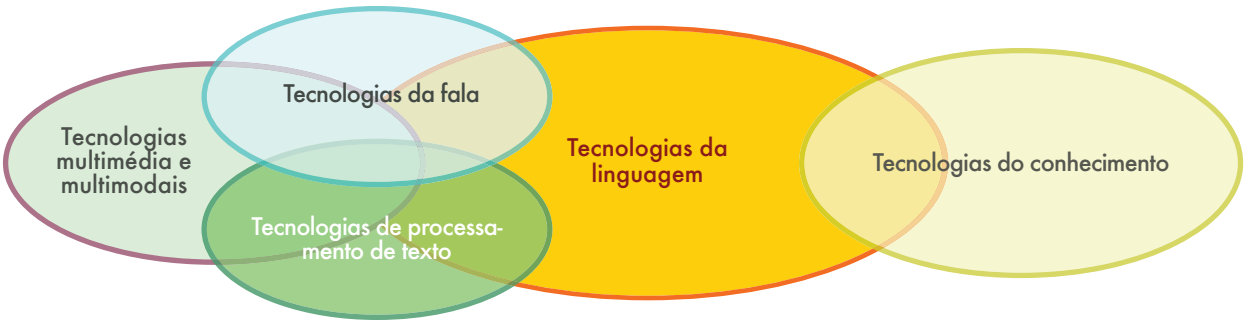
A tecnologia da linguagem constitui uma área de investigação autónoma com uma vasta literatura. Para uma introdução, o leitor interessado poderá consultar as seguintes referências [29, 30, 31, 32].

Em preparação da discussão sobre as áreas de aplicação nucleares apontadas acima, descrever-se-á brevemente a arquitetura típica de um sistema de tecnologia da linguagem.

4.1 ARQUITETURAS DE APLICAÇÕES

As aplicações mais usuais para o processamento da linguagem são constituídas por vários componentes que refletem diferentes aspetos da linguagem. A Figura 2 mostra, de um modo bastante simplificado, a arquitetura que pode ser encontrada num sistema típico de processamento de texto. Os três primeiros módulos ocupam-se da estrutura e do significado do texto de entrada:

1. pré-processamento: limpeza dos dados, análise ou remoção da formatação, e deteção do idioma, etc;



1: A tecnologia da linguagem em contexto

2. análise gramatical: deteção do verbo e dos seus complementos e modificadores, deteção de elementos de outras categorias, identificação da estrutura das frases;

3. análise semântica: desambiguação (por exemplo, qual das aceções de *bateria* é a usada em determinado contexto?), resolução de anáforas (por exemplo, que pronome recupera a referência de que outra expressão na frase?), e representação do significado da frase num modelo interpretável pela máquina.

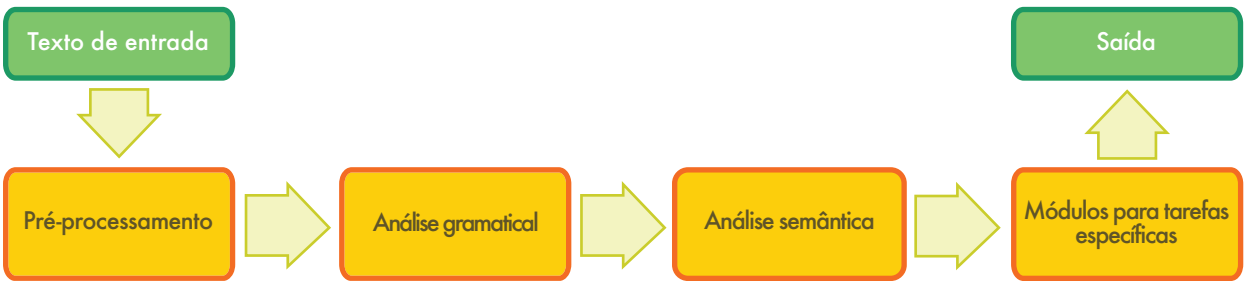
Após a análise do texto, alguns módulos específicos podem executar outro tipo de operações, como a sumarização automática ou uma busca em bases de dados, entre outras.

4.2 ÁREAS CENTRAIS DE APLICAÇÃO

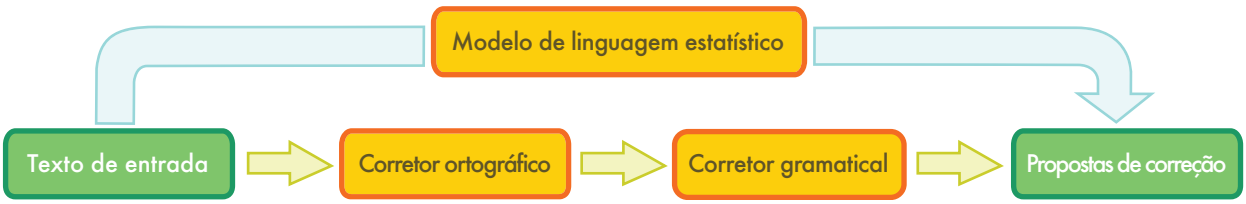
Apresentar-se-ão, em seguida, algumas aplicações centrais na área da tecnologia da linguagem: verificação de linguagem, busca na web, tecnologia da fala e tradução automática.

4.2.1 Verificação da Linguagem

Quem tiver usado uma ferramenta de processamento de texto, como o MS Word, sabe que esta tem um corretor ortográfico que destaca possíveis erros ortográficos e propõe correções. Os primeiros programas de verificação ortográfica comparavam uma lista de palavras extraídas do texto a analisar com o que constava de um dicionário com palavras corretamente escritas. Hoje em



2: Arquitetura típica de uma aplicação para o processamento de texto



3: Corretor ortográfico e sintático: modelo estatístico (em cima) e modelo baseado em regras (em baixo)

dia, esses programas tornaram-se bem mais sofisticados. Além de usarem algoritmos para a análise de texto afinados para a linguagem em apreço, detetam erros relacionados com a morfologia (por exemplo, formação do plural) e a sintaxe, tais como a ausência de um verbo ou a falta de concordância com o sujeito em pessoa e número (por exemplo, como em *elas *escreve uma carta*), etc. Ainda assim, a maioria dos corretores ortográficos não alertará para um potencial erro na segunda destas duas frases:

Fizemos jogos tradicionais, incluindo o *jogo do pião*.
Fizemos jogos tradicionais, incluindo o *jogo do peão*.

Para lidar com este tipo de erros, é necessária a formulação de regras gramaticais específicas da língua (o que implica um elevado grau de especialização e trabalho manual) ou o uso de um modelo de linguagem estatístico, como ilustrado na Figura 3. Este tipo de modelo calcula a probabilidade de uma determinada palavra ocorrer num determinado contexto. Para o exemplo acima referido, *o jogo do pião* é uma sequência de palavras muito mais provável do que o *jogo do peão*. Um modelo estatístico pode ser automaticamente obtido recorrendo-se a uma grande quantidade de dados da língua, que se costuma designar por um corpus.

A verificação da linguagem não se limita aos processadores de texto. É também usada em sistemas de apoio ao autor (*authoring support systems*). Estes sistemas são aplicações que apoiam a redação de manuais e outra documentação para as áreas das tecnologias da informação complexas, cuidados de saúde ou engenharia, entre outros. Temendo as reclamações dos clientes devido à utilização errada dos produtos ou devido aos danos resultantes de uma possível má interpretação dos manuais de instrução, as empresas prestam cada vez mais atenção à qualidade técnica da documentação quando se dirigem ao mercado internacional. Os avanços na área da tecnologia da linguagem levaram ao desenvolvimento de aplicações de apoio à elaboração de textos, que auxiliam o redator de documentação técnica no uso de vocabulário e de estruturas de frases, de acordo com certas regras e restrições terminológicas.

O uso de corretores ortográficos não se limita aos processadores de texto. Também se aplica a sistemas de apoio aos autores de textos especializados.

Para além do corretor ortográfico associado ao MS Word, existem outras ferramentas de correção ortográfica para o português. Em Portugal, é comercializado o FLIP, um software que disponibiliza vários produtos na área da verificação ortográfica e sintática para o português europeu e do Brasil. O CoGrOO, para o Open Office, é um corretor gramatical para o português do Brasil. Também para esta variedade do português, e partindo de um algoritmo concebido pelo Instituto de Computação da Universidade Estadual de Campinas (UNICAMP), o Núcleo Interinstitucional de Lingüística Computacional (NILC) desenvolveu o corretor Re-

Gra, que é parte integrante do MS Word e do processador de texto REDATOR.

Além dos corretores ortográficos e dos sistemas de apoio ao autor, este tipo de verificação da língua é também importante na área da aprendizagem de línguas assistida por computador e nas aplicações de correção automática de pesquisas enviadas para motores de busca da internet, como é o caso das sugestões do Google "Será que quis dizer ...".

4.2.2 Busca na Web

A busca na web, em intranets ou em bibliotecas digitais é provavelmente a tecnologia da linguagem mais utilizada mas também a menos desenvolvida nos dias de hoje. Na Figura 4 encontra-se uma representação esquemática dos seus principais componentes.

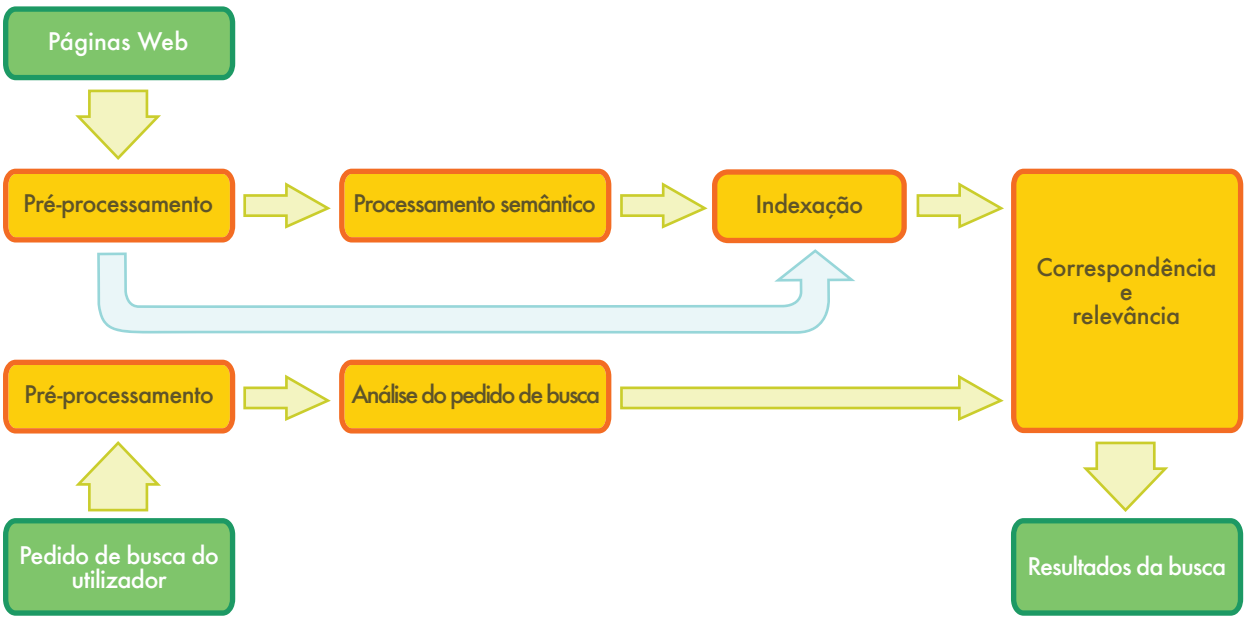
O motor de busca Google, surgido em 1998, recebe atualmente cerca de 91% dos pedidos de busca que se fazem na web em todo o mundo [33]. O verbo *googlar* passou a ter uma entrada no dicionário de Português online da Porto Editora [34]. Nem a interface de busca nem a apresentação dos resultados obtidos sofreram alterações significativas desde a primeira versão deste motor de busca. Na versão atual, o Google oferece correção ortográfica para as palavras com erros ortográficos. A sua capacidade de busca semântica, que desde 2009 se encontra incorporada no seu algoritmo, permite-lhe melhorar a precisão dos resultados através da análise do significado dos termos do pedido de busca no seu contexto [35].

A história de sucesso do Google mostra que, na posse de um grande volume de dados e de técnicas de indexação eficiente de dados, uma abordagem essencialmente baseada em estatística pode levar a resultados satisfatórios.

No entanto, para uma busca de informação mais elaborada, é essencial integrar conhecimentos linguísticos mais profundos. Experiências realizadas em laboratório, com recurso a thesauri e bases de dados ontológicas (como a ontologia lexical WordNet), têm apresentado avanços ao permitir que se encontre uma página com base nos sinónimos dos termos da busca (por exemplo, para uma busca por *energia atómica*, busca-se automaticamente também por *energia nuclear* e *centrais nucleares*, etc). Neste contexto, para o português (europeu ou do Brasil), será útil a ontologia lexical MultiWordnet.PT [36], para o português europeu, a WordNet.PT [37], e para o português do Brasil, o Thesaurus Eletrônico para o Português (TEP), em desenvolvimento como parte do projeto WordNet.BR.

A próxima geração de motores de busca terá de incluir tecnologia da linguagem muito mais sofisticada. Se em vez de uma lista de palavras-chave, a busca consistir numa pergunta ou noutro tipo de frase, a obtenção de respostas relevantes para esta consulta vai requerer não só uma análise da frase a nível sintático e semântico, como também a disponibilização de uma indexação que permita uma recuperação rápida dos documentos pertinentes. Suponhamos, por exemplo, que um utilizador introduz a seguinte busca: *Quais são as empresas que foram compradas por outras empresas nos últimos cinco anos?* Para se alcançar uma resposta satisfatória, é necessário proceder-se a uma análise gramatical da frase para obter a sua estrutura e determinar que o utilizador está à procura de empresas que foram compradas e não de empresas que compraram outras; é igualmente preciso processar a expressão *últimos cinco anos* para descobrir quais os anos a que ela se refere exatamente, etc.

Adicionalmente, é necessário que o pedido de busca seja comparado com uma grande quantidade de dados não estruturados, com o objetivo de encontrar parte (ou partes) da informação de que o utilizador está à procura. Este processo é normalmente referido como recuperação de informação (*information retrieval*) e envolve tarefas de busca em documentos considerados relevantes. No caso da busca acima referida, para se obter



4: Arquitectura da busca na web

uma lista de empresas é ainda necessário extrair a informação de que uma dada sequência de palavras num documento se refere ao nome da empresa. Esta tarefa é realizada através de ferramentas que executam aquilo que na área se designa por reconhecimento de expressões nomeadoras de entidades (*named entity recognition*).

A próxima geração de motores de busca terá de incluir a tecnologia da linguagem com um grau muito mais elevado de sofisticação.

Mais exigente ainda é fazer uma busca por documentos escritos em línguas diferentes do idioma dos termos de busca. Para a recuperação de informação transversal a diferentes línguas, há que traduzir automaticamente a busca para todas as línguas alvo possíveis e transferir a informação recolhida de volta para a língua fonte.

Face à crescente percentagem de dados disponíveis em formatos não textuais, há uma necessidade de serviços que permitam a recuperação de informação multimédia, ou seja, a busca de informação em imagens, em áudio e em vídeo. Para ficheiros de áudio e vídeo, esta tarefa envolve um módulo de reconhecimento da fala que tem por função converter a fala em formato textual ou numa representação fonética em relação aos quais se possa estabelecer uma correspondência com as buscas que os utilizadores possam fazer.

No final dos anos 90, começaram a ser desenvolvidos em Portugal vários motores de busca. O AEIOU surgiu em 1996 e foi posteriormente comprado pelo grupo Impresa, sendo transformado num portal de conteúdos [38]. O Sapo foi lançado em 1997 como motor de busca, tornando-se mais tarde um portal e sendo agora um fornecedor de serviços de internet propriedade da PT Multimédia [39]. Foram também criadas versões deste motor de busca para Angola, Cabo Verde, Moçambique e Timor-Leste. Hoje em dia, embora tenham sido criados muitos outros motores de busca em Portugal (Busca Online, Clix, Guianet, Netindex, entre outros) [40], são poucas as empresas portuguesas que con-

tinuam a fornecer serviços autónomos de busca, sendo o Google.pt tido como o mais popular.

No Brasil encontram-se exemplos de motores de busca direcionados apenas para sites brasileiros – como o Achei [41] ou o Giga Busca [42] –, sendo a sua cobertura e o seu alcance limitados. Há que destacar o motor de busca METAMINER, desenvolvido em 1996 pela Universidade Federal de Minas Gerais, mais tarde integrado no portal UOL. O Google.br é por isso tido como o motor de busca dominante no Brasil.

4.2.3 Interação por Fala

A interação através de fala é uma das muitas áreas de aplicação que dependem da tecnologia da fala, ou seja de tecnologia que processa os sons da linguagem. A tecnologia da fala é usada para criar interfaces que permitem ao utilizador interagir com máquinas usando linguagem falada em vez de, por exemplo, um monitor, um teclado ou um rato. Atualmente estas interfaces com o utilizador baseadas em voz podem ser parcial ou totalmente automatizadas e são geralmente utilizadas por empresas para oferecerem serviços por telefone aos seus clientes, empregados ou associados. Os negócios na área da banca, logística, transportes públicos ou telecomunicações são dos que mais fortemente apostam neste tipo de aplicações. A tecnologia da fala proporciona ainda outros tipos de utilizações, nomeadamente interfaces para certos dispositivos, como por exemplo, os sistemas de navegação presentes nos carros, ou o recurso à linguagem oral como alternativa às modalidades de input/output existentes em interfaces gráficas, como acontece com os smartphones.

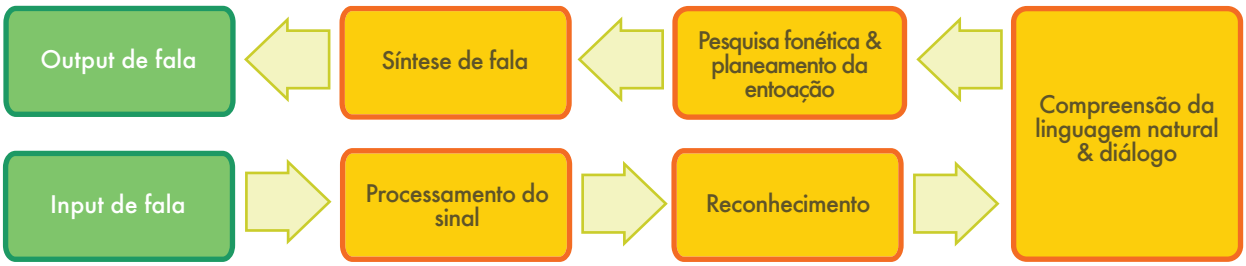
A tecnologia da fala é a base para se criar interfaces que permitem ao utilizador interagir com máquinas usando a voz em vez de um teclado ou um rato.

Como ilustrado na Figura 5, sobre sistemas de diálogo baseados em voz, a tecnologia da fala compreende três dimensões principais:

1. O **reconhecimento automático da fala** determina que palavras foram efetivamente proferidas numa sequência de sons produzidos por um utilizador.
2. A **gestão do diálogo** determina que ação deve ser realizada tendo em conta o input do utilizador e a funcionalidade do próprio sistema.
3. A **síntese de voz** (texto-para-fala) transforma o output do sistema em sons para o utilizador.

Um dos grandes desafios dos sistemas de reconhecimento automático da fala consiste em reconhecer com precisão as palavras proferidas por um utilizador. Isto pode implicar restringir-se o leque de enunciados admissíveis a um conjunto limitado de palavras-chave, ou proceder-se à criação manual de modelos de linguagem que cubram uma grande variedade de enunciados em linguagem natural. Através da utilização de técnicas de aprendizagem automática, os modelos de linguagem podem também ser gerados automaticamente a partir de corpora de fala, ou seja, de grandes coleções de ficheiros áudio com fala e respetivas transcrições textuais. Restringir-se o leque de enunciados admissíveis força porém as pessoas a utilizarem a interface de voz de uma forma rígida e reduz a sua aceitação por parte dos utilizadores. Interfaces de tipo alternativo, que recorrem a modelos de linguagem e permitem ao utilizador expressar a sua intenção de forma mais flexível – numa interação desencadeada, por exemplo, pela pergunta "*Como posso ajudá-lo?*" –, têm por isso uma melhor aceitação. Mas esta alternativa envolve a criação, afinação e manutenção de modelos de linguagem, o que pode fazer aumentar os custos de modo muito significativo.

Os sistemas de reconhecimento do português europeu e do português do Brasil têm um bom desempenho em



5: Sistema de diálogo baseado em voz

geral, obtendo resultados de reconhecimento moderadamente bons, e têm sido mantidos de forma ativa. A grande maioria destes sistemas não se encontra disponibilizada de forma livre e os sistemas desenvolvidos nos laboratórios, em particular, não apresentam conformidade com padrões estabelecidos. Alguns sistemas usam grandes vocabulários, para transcreverem notícias, por exemplo. Alguns são específicos para um certo domínio, usando um vocabulário limitado (para tarefas circunscritas, e. g. na área da medicina), sendo a adaptação a um novo domínio possível com recursos apropriados.

As empresas tendem a usar enunciados pré-gravados por locutores profissionais para gerar o output de uma interface de voz. Para enunciados estáticos em que a formulação não depende de contextos particulares nem de dados pessoais do utilizador, isto permitirá uma experiência do utilizador satisfatória. No entanto, quanto mais dinâmico for o conteúdo de um enunciado que o sintetizador tem de produzir mais hipóteses há de os resultados de output apresentarem uma prosódia pobre, resultante da mera concatenação de pedaços de áudio. Recorrendo-se a técnicas de otimização, os atuais sistemas de texto-para-fala têm apresentado cada vez melhores resultados na produção de enunciados dinâmicos que soam com naturalidade.

O estado da arte da síntese de fala para o português é similar ao do reconhecimento de fala. Poucos sistemas são acessíveis de forma livre e os dados de fala necessários para criar uma voz não se encontram disponíveis. No entanto, a maturidade dos sistemas de síntese para uso generalizado parece ainda assim ser maior em várias aplicações: dispositivos GPS, centros de atendimento telefónico, avatares, websites, etc.

A última década tem sido caracterizada por uma padronização das interfaces de interação por fala em termos dos seus vários componentes tecnológicos. Houve também uma forte consolidação do mercado nos últimos dez anos, em particular nas áreas de reconhecimento e síntese da fala. Os mercados nacionais dos países do G20 são dominados por apenas cinco atores globais, sendo a Nuance (EUA) e a Loquendo (Itália) as empresas mais proeminentes. Em 2011, a Nuance anunciou a aquisição da Loquendo, o que representa mais um passo na consolidação do mercado.

No mercado português de texto-para-fala, existem algumas pequenas empresas, como a SVOX e a Voice Interaction, procurando esta última diferenciar-se por disponibilizar vozes não apenas para o português europeu e do Brasil, mas também para as variedades africanas do português. No mercado brasileiro a empresa VOCALISE oferece produtos e serviços nesta área (texto-para-fala, fala-para-texto, reconhecimento automático de fala, busca em fala gravada, etc), com a particularidade de estar muito próxima das grandes universidades da zona de São Paulo e Campinas [43]. É de destacar também o número crescente de empre-

sas estrangeiras que se estabelecem junto das universidades e que têm demonstrado interesse nas diferentes variedades do português do Brasil.

No que respeita à tecnologia e know how para gestão de diálogo, a DigA é a única aplicação completa construída especificamente para o português europeu: é de domínio público mas não está disponível em código aberto. A aplicação Olympus SDS, de código aberto, foi adaptada com sucesso para o português mas ainda não foi amplamente testada. Dos vários módulos exigidos por sistemas de diálogo, o gestor de diálogo é o único módulo que pode ser usado para qualquer língua. Os outros módulos existem embora não sejam usualmente de livre acesso nem estejam disponíveis em código aberto.

Olhando para o futuro, anteveem-se mudanças significativas devido à disseminação dos smartphones enquanto nova plataforma para a gestão de relações com clientes, em acumulação com o telefone fixo, a internet e o correio eletrónico. Isto afetará também a forma como a tecnologia da fala é usada. A longo prazo, haverá menos interfaces baseadas em voz para serem usadas por telefone e a utilização da linguagem falada desempenhará um papel cada vez maior enquanto input amigável para smartphones. Esta tendência será impulsionada pelas melhorias graduais, que se irão obtendo no futuro próximo, em termos da precisão do reconhecimento de fala independente do falante feito através serviços de ditado, serviços esses que são já oferecidos como serviços centralizados para utilizadores de smartphones.

Para o português europeu, tem havido recentemente investigação dirigida para novas aplicações, nomeadamente nas áreas da saúde e do ensino da língua. Alguns projetos procuram, por exemplo, desenvolver e testar ferramentas para apoiar o ensino da pronúncia ou para jogos "sérios" para a aquisição de vocabulário e da gramática. No caso da saúde, decorrem projetos que estudam a fala dos idosos e o seu impacto no desempenho das ferramentas de reconhecimento da fala, com vista a ajudar a recuperação de doentes com perturbações da fala, como a afasia.

4.2.4 Tradução Automática

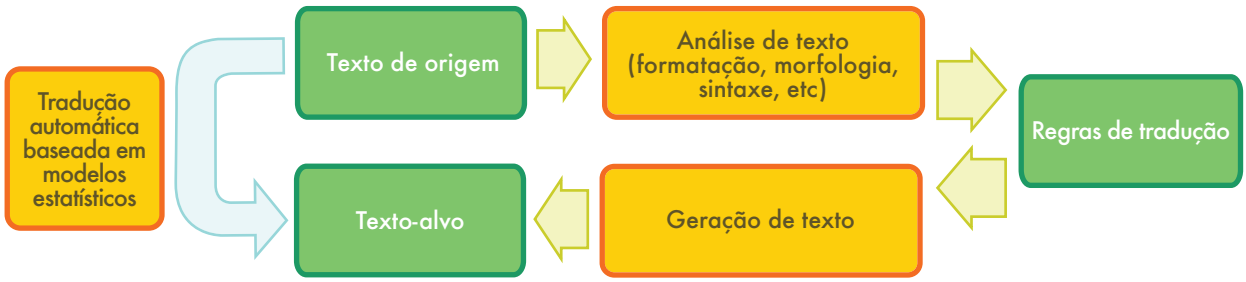
A ideia de usar computadores para a tradução das línguas naturais surgiu em 1946 e veio a merecer financiamentos substanciais nos anos 50 e novamente nos anos 80. A tradução automática encontra-se longe de corresponder, porém, às expectativas que gerou nos primeiros anos de investigação.

No seu nível mais básico, a tradução automática pode ser realizada através de uma mera substituição das palavras de uma língua por palavras de outra língua. Isto poderá ser útil em domínios com terminologias restritas e que façam uso de uma linguagem controlada, como por exemplo, os boletins meteorológicos. Contudo, para uma boa tradução de textos menos padronizados, é necessário fazer corresponder as unidades de texto maiores (sintagmas, frases ou mesmo textos completos) às suas contrapartes mais próximas na língua alvo. Neste caso, a maior dificuldade reside no facto de a linguagem humana ser ambígua. A desambiguação de palavras apresenta um enorme desafio a vários níveis. Por exemplo, a nível lexical, *banco* apresenta pelo menos duas aceções, "peça de mobiliário" ou "instituição financeira", o que é ilustrado no seguinte exemplo:

O Pedro viu a rapariga no banco.

Dependendo do contexto em que ocorra, esta frase tanto pode indicar que o Pedro viu a rapariga na instituição bancária ou no assento.

A ambiguidade sintática também apresenta grandes desafios, como é ilustrado pelos dois exemplos abaixo. Repare-se que as frases são estruturalmente idênticas, mas na primeira o sintagma preposicional introduzido por *com* causa ambiguidade, e na segunda não – o telescópio foi usado pelo Pedro para ver a rapariga, ou a rapariga usava o telescópio quando foi vista pelo Pedro:



6: Tradução Automática: modelo estatístico (esquerda) e modelo baseado em regras (direita)

O Pedro viu a rapariga com o telescópio.
O Pedro viu a rapariga com o boné.

Uma forma de construir sistemas de tradução automática consiste em usar regras linguísticas. Para traduções entre línguas aproximadas, a tradução direta palavra a palavra pode ser útil. Mas os sistemas mais sofisticados são baseados em regras e em conhecimento linguístico que ajudam a analisar o texto de entrada e a criar uma sua representação intermédia a partir da qual geram o texto da língua alvo. O sucesso destes métodos está fortemente dependente da disponibilidade não só de grandes léxicos – com informação morfológica, sintática e semântica –, como também de grandes conjuntos de regras gramaticais concebidas cuidadosamente por linguistas especializados. Alguns dos mais importantes sistemas de tradução automática baseados em regras, como o LOGOS, o Apertium ou o SYSTRAN, estão disponíveis para a língua portuguesa.

A partir dos finais dos anos 80, quando os recursos computacionais se tornaram mais baratos, começou a surgir um maior interesse na criação de modelos estatísticos para a tradução automática. Os parâmetros destes modelos derivam da análise de corpora bilingues, como por exemplo, o corpus paralelo Europarl, que contém as atas do Parlamento Europeu em 21 línguas diferentes. Com um volume de dados suficiente, através do processamento de versões paralelas e da busca por padrões prováveis de palavras, a tradução automática baseada em estatística funciona suficientemente bem para produzir uma tradução aproximada na língua alvo. Além da vantagem de ser necessário um menor esforço humano, a tradução automática baseada em estatística pode também cobrir particularidades da língua de que os outros sistemas não dão conta, como é o caso, por exemplo, das expressões idiomáticas. Contudo, ao contrário dos sistemas baseados em regras linguísticas, este tipo de abordagem tende a gerar, muitas mais vezes, resultados com erros gramaticais.

Adicionalmente, e no caso do português em particular, a falta de recursos para a desambiguação de aceções de palavras – dados (ontologias lexicais e corpora anotados) e software desenvolvido a partir desses dados – é uma das razões para que os resultados dos sistemas de tradução automática existentes sejam ainda mais insatisfatórios.

A Figura 6 sintetiza diagramaticamente estas duas abordagens para a tradução automática, baseada em regras e baseada em estatística. Devido ao facto de os pontos fortes e os pontos fracos destes dois tipos de abordagem para a tradução automática serem complementares, os investigadores têm-se concentrado em aperfeiçoar abordagens híbridas, combinando ambas as metodologias. Uma das formas de pôr em prática esta ideia consiste em utilizar tanto o modelo baseado em regras como o modelo baseado em estatística e ter um módulo de seleção que decida o melhor output para cada frase. No entanto,

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

7: Tradução automática entre 22 línguas oficiais da UE – Machine translation between 22 EU-languages [44]

para frases mais longas, por exemplo, com mais de doze palavras, os resultados estão longe de serem perfeitos.

Apesar de haver uma investigação significativa nesta área da tecnologia, os sistemas híbridos têm sido, até agora, menos bem sucedidos em termos comerciais do que em termos de investigação.

Há ainda um grande potencial para se melhorar a qualidade dos sistemas de tradução automática. De entre os desafios existentes, destacam-se a adaptação dos recursos linguísticos a domínios ou áreas de utilização específicos, e a sua integração em sistemas que já têm bases de dados terminológicas e memórias para tradução. Além disso, a maioria dos atuais sistemas é direcionada para o inglês, havendo poucos sistemas para a tradução entre pares de línguas de e para o português.

A qualidade dos sistemas de tradução automática costuma ser avaliada através de campanhas de avaliação, que permitem a comparação do desempenho dos sistemas perante diferentes metodologias e diferentes línguas. O quadro da Figura 7 foi preparado no âmbito do projeto Euromatrix+, apoiado pela Comissão Europeia. Mostra o resultado de uma campanha de avaliação para o desempenho de um mesmo sistema de tradução automática baseado em estatística, o MOSES, na tradução entre os pares de línguas obtidos para 22 das 23 línguas oficiais da União Europeia (com exceção do irlandês). Os resultados estão ordenados de acordo com a classificação BLEU, que atribui as pontuações mais elevadas às melhores traduções [45]. Um tradutor humano conseguirá, em regra, uma avaliação de cerca de 80 pontos.

Os melhores resultados (a azul e a verde) foram obtidos tanto para línguas que têm beneficiado de consideráveis esforços de investigação, apoiados por programas de financiamento à Investigação e Desenvolvimento, como da existência de corpora paralelos – como é o caso, por exemplo, das línguas inglesa, francesa, neerlandesa, es-

panhola ou alemã. Os piores resultados (a vermelho) dizem respeito a línguas que não beneficiaram de esforços semelhantes ou que estão em pares de tradução com línguas de famílias linguísticas muito diferentes.

4.3 OUTRAS ÁREAS DE APLICAÇÃO

A construção de aplicações na área da tecnologia da linguagem envolve uma série de tarefas que nem sempre são diretamente percetíveis ao nível da interação com o utilizador mas que asseguram funcionalidades significativas nos "bastidores" dos sistemas em questão. Essas tarefas e suas funcionalidades têm constituído tópicos cruciais de investigação, tendo-se tornado subáreas autónomas da tecnologia da linguagem.

As aplicações de tecnologia da linguagem asseguram funcionalidades-chave nos "bastidores" de sistemas mais amplos.

Os sistemas de resposta a perguntas, por exemplo, tornaram-se numa das áreas de investigação mais ativas, tendo levado à construção de corpora anotados e ao estabelecimento de competições científicas específicas. O objetivo é passar de uma busca baseada em palavras-chave (à qual o motor de busca deve responder com um conjunto de documentos potencialmente relevantes) para o cenário em que o utilizador coloca uma questão concreta e o sistema produz uma única resposta, como no seguinte exemplo:

> Pergunta: *Com que idade Neil Armstrong pisou a Lua?*
>
> Resposta: *38 anos.*

Estando esta área relacionada com o que foi acima referido sobre a busca na web, ela tem porém agrupado uma série de tópicos de investigação específicos, como por exemplo: que tipos de perguntas existem e como é que devem ser tratados; como é que os documentos que podem conter a resposta devem ser analisados e comparados (será que fornecem respostas contraditórias?); que nível de confiança atribuir a uma informação específica extraída (a resposta) levando em consideração o contexto, etc.

As questões acima colocadas estão, por sua vez, relacionadas com a tarefa de extração de informação, uma área que foi muito popular e influente no deslocamento espistemológico do início dos anos 90 em direção à exploração de métodos estatísticos.

A extração de informação tem como objetivo identificar conteúdos específicos de informação em determinado tipos de documentos. Por exemplo, pode consistir em identificar os agentes principais na aquisição de uma dada empresa, tal como esta aquisição é relatada nos jornais. Uma outra aplicação, por exemplo, diz respeito a relatórios sobre incidentes terroristas, em que o objetivo consiste no mapeamento de partes de textos em partes de uma ficha de informação (*information template*) que registam, por exemplo, a informação sobre o agressor, o alvo, a hora, o local e os resultados do incidente. O preenchimento de fichas de informação relativas a domínios específicos é pois a característica central da extração de informação, o que faz dela mais um caso de tecnologia da linguagem a funcionar nos "bastidores" e uma das subáreas da tecnologia da linguagem.

A sumarização e a geração automática de textos, por sua vez, constituem outras duas áreas que podem desempenhar um papel de tecnologia de apoio nos "bastidores" ou podem funcionar como aplicações individualizadas. A sumarização consiste na tarefa de fornecer o que é essencial num texto numa sua versão mais reduzida, sendo uma das funcionalidades disponíveis, por exemplo, no MS Word. Esta aplicação funciona sobretudo com base em métodos estatísticos: identifica

primeiramente palavras "importantes" num texto (que podem ser, por exemplo, aquelas que apresentam uma frequência elevada nesse texto mas que são muito menos frequentes no uso geral que os falantes fazem da língua) e em seguida seleciona as frases que contêm essas palavras "importantes". Estas frases são então marcadas no documento, ou extraídas, e é a partir delas que se irá construir o resumo. Neste cenário, que é de longe o mais aplicado, a sumarização corresponde ao processo de extração de frases: o texto é reduzido a um subconjunto das suas frases. Todas as aplicações comerciais de sumarização automática de textos funcionam deste modo.

Uma abordagem alternativa, que tem estado a ser investigada, consiste em sintetizar efetivamente frases novas que não ocorrem no texto de origem. Esta tarefa exige uma compreensão mais aprofundada do texto e por isso tem permitido até agora soluções menos robustas. Cabe notar que um gerador automático de texto deste género não representa, em regra, uma aplicação individual, encontrando-se embebido numa aplicação mais vasta, como é o caso dos sistemas de informação hospitalares, nos quais os dados dos doentes são recolhidos, armazenados e processados. A geração automática de relatórios será apenas uma das suas muitas funcionalidades.

Nestas áreas, a investigação tem recaído muito menos sobre a língua portuguesa do que sobre outras línguas, sobretudo a língua inglesa, em relação à qual sistemas de resposta a perguntas, de extração de informação e de sumarização automática têm sido objeto, desde a década de 90, de inúmeros concursos para atribuição de financiamento à Investigação e Desenvolvimento, como os organizados pela DARPA/NIST, nos Estados Unidos. Este apoio tem contribuído significativamente para o avanço do estado da arte em tecnologia da linguagem, focado porém no inglês.

A língua portuguesa, tal como muitas outras línguas, não tem recebido apoio suficiente para poder ser processada ao nível do estado da arte, e muito menos para que o seu estudo possa oferecer uma maior contribuição para o avanço da fronteira do conhecimento neste domínio científico e tecnológico.

A investigação e as aplicações desenvolvidas estão esmagadoramente direcionadas para o inglês. Sendo os resultados iniciais obtidos para o português promissores, a investigação referente à lingua portuguesa carece de um impulso decidido para ser continuada e aprofundada.

Nos laboratórios de investigação foram desenvolvidos protótipos de sistemas de resposta a perguntas para o português, como por exemplo o XisQuê [46], da Universidade de Lisboa, que procura as respostas para as perguntas na web dos textos em língua portuguesa (disponível para demonstração em http://xisque.di.fc.ul.pt). Sendo os resultados promissores, a investigação referente à lingua portuguesa carece porém de ser continuada e aprofundada.

Quanto aos sistemas de sumarização automática, aqueles que utilizam apenas métodos estatísticos são, em grande medida, independentes da língua e neste caso, encontram-se disponíveis alguns protótipos de sumarizadores para o português, como por exemplo, o GistSum, da Universidade de São Paulo.

No que respeita à geração automática de texto, existem componentes reutilizáveis cujo uso tem sido tradicionalmente limitado à construção de módulos que geram estruturas de superfície (as gramáticas de geração). Mas também aqui as aplicações desenvolvidas estão esmagadoramente direcionadas para o inglês, não havendo nesta área ferramentas disponíveis para o português.

4.4 FORMAÇÃO ACADÉMICA

A tecnologia da linguagem é uma área altamente interdisciplinar que envolve a combinação das competências

de informáticos, linguistas, matemáticos, filósofos e psicolinguistas, entre outros.

Em Portugal, a área da tecnologia da linguagem tem vindo a ser promovida em várias universidades quer em termos de investigação quer em termos educativos, em cursos de licenciatura, mestrado e doutoramento. No Ensino Superior há uma oferta razoável nesta área, encontrando-se as disciplinas relevantes integradas em cursos oferecidos por Departamentos de Informática ou de Ciências da Linguagem.

Na Universidade de Lisboa, a par de diversas disciplinas em diferentes níveis de ensino, (incluídas num minor em Processamento de Linguagem Natural, no mestrado e no doutoramento em Engenharia Informática e nos programas de mestrado e doutoramento em Ciência Cognitiva), existem centros de investigação dedicados à tecnologia da linguagem. O Departamento de Informática, da Faculdade de Ciências, acolhe uma unidade dedicada ao processamento computacional do português (o grupo NLX), que entre várias outras atividades, assegura o LX-Center [47], um centro online de serviços de processamento linguístico e de demonstração da tecnologia da linguagem, e coordena um dos quatro projectos europeus da Rede de Excelência META-NET. O Centro de Linguística (CLUL), da Faculdade de Letras, conta com uma longa tradição na produção de recursos linguísticos – quer a nível do português padrão, quer a nível dialetal ou mesmo da história da língua –, tendo construído um corpus de grande escala, de que resultou o desenvolvimento de outros recursos mais específicos, disponíveis online.

O Instituto Superior Técnico (IST), em Lisboa, além de oferecer cursos em tecnologia da linguagem, também assegura um programa de doutoramento em Ciências da Computação em colaboração com outras universidades portuguesas e com a Carnegie Mellon University. O INESC-ID é uma instituição de investigação associada ao IST e o seu Laboratório de Sistemas de Língua Falada (L2f) é um centro líder na produção de sistemas de reconhecimento e síntese da fala.

A Universidade Nova de Lisboa tem também cursos e unidades de investigação activas neste campo da tecnologia da linguagem, nomeadamente o Centro de Investigação em Tecnologias de Informação (CITI) e o Centro de Linguística (CLUNL).

Ainda em Lisboa, existe o Instituto de Linguística Teórica e Computacional (ILTEC), que foi criado para albergar o projecto EUROTRA.

Na Universidade do Porto, dois centros têm feito trabalho em ciência e tecnologia da linguagem natural, nomeadamente o Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC) e o Centro de Linguística (CLUP).

A actividade neste campo de forma alguma se restringe às duas maiores cidades, Lisboa e Porto. No resto do país, existem várias outras universidades que oferecem também cursos na área da ciência e tecnologia da linguagem e que acolhem centros de investigação.

É o caso do Centro de Investigação em Tecnologias da Informação (CITI-UE), na Universidade de Évora.

Na Universidade de Coimbra, destacam-se o Centro de Estudos de Linguística Geral e Aplicada (CELGA) e o Instituto de Telecomunicações (IT).

Cabe indicar igualmente o Centro de Tecnologia da Linguagem Humana e Bioinformática (HULTIG), na Universidade da Beira Interior, assim como o Centro de Estudos Humanísticos (CEHUM), na Universidade do Minho.

A Universidade do Algarve tem cooperado com o programa europeu Erasmus na realização de um mestrado na área do Processamento de Linguagem Natural.

A tecnologia da linguagem tem vindo a ser promovida em várias universidades quer em termos de investigação que em termos educacionais.

No Brasil, tem-se assistido igualmente a uma atividade considerável na área da tecnologia da linguagem, tanto no ensino como na investigação, que se concentra sobretudo nas áreas Sul e Sudeste do país, com particular destaque para as áreas urbanas de São Paulo, Porto Alegre e Rio de Janeiro. Os cursos têm sido ministrados mais a nível de pós-graduações (mestrados e doutoramentos) do que de licenciatura. Recentemente, foi elaborado o Programa Nacional de Pós-Graduação 2011-2020, com que se procura reforçar o interesse pela investigação inter e multidisciplinar.

Nos outros países de língua portuguesa, a área da tecnologia da linguagem apresenta pouco ou nenhum desenvolvimento, sendo que a recolha de dados e o desenvolvimento de recursos e ferramentas orientados para as outras variedades do português têm sido realizados principalmente pelos centros de investigação em Portugal.

4.5 PROJETOS E INICIATIVAS

Em Portugal, a atividade na área da tecnologia da linguagem tem sido sustentada por iniciativas, projetos e programas de investigação levados a cabo nas últimas décadas. Para efeitos ilustrativos, nesta seção referiremos apenas alguns.

Um dos primeiros e mais importantes programas nesta área foi o EUROTRA, um ambicioso programa sobre tradução automática criado e financiado pela Comissão Europeia desde o final dos anos 70 até 1994. Portugal entrou neste programa em 1986 através do ILTEC, criado especificamente para este propósito e contando com investigadores sobretudo das Universidades de Lisboa e do Porto. Este programa teve um impacto duradouro a nível europeu. Constituiu um impulso decisivo para a prossecução de atividades no âmbito da tecnologia da linguagem em Portugal e para o surgimento e consolidação de uma comunidade de investigadores nesta área no país. O projeto LE-PAROLE, desenvolvido no final dos anos 90, com a participação do CLUL e do INESC-ID, foi outro projeto-chave europeu na área da tecnologia da linguagem que envolveu a língua portuguesa. Dos seus resultados, destaca-se a construção de corpora e léxicos de acordo com modelos integrados de constituição e descrição de materiais, o que permite estabelecer ligações multilingues e dar apoio a um grande número de aplicações. Para cada língua, foi construído um corpus de 20 milhões de palavras, comparável no que respeita à composição e codificação, que incluiu um subcorpus anotado de 250 mil palavras. Foi também constituído um léxico para cada língua, incluindo o português, composto por 20 mil entradas, com informação sintática e morfológica.

Parte deste corpus foi alargado e enriquecido no projeto TagShare, levado a efeito na Universidade de Lisboa pelo Departamento de Informática (NLX) e pelo Centro de Linguística (CLUL), em 2005. Este projeto desenvolveu um conjunto de recursos linguísticos e de ferramentas que permitem melhorar o processamento computacional do português. Obteve-se um corpus de 1 milhão de palavras linguisticamente anotadas e manualmente revistas por especialistas – o corpus CINTIL [48] –, assim como todo um conjunto de ferramentas para segmentação, anotação de categoria morfossintática, flexão, lematização, reconhecimento de unidades lexicais multipalavra, reconhecimento de expressões nomeadoras de entidades, etc. Os esquemas de anotação desenvolvidos no âmbito deste projecto tornaram-se num padrão de facto para o português no campo da tecnologia da linguagem, sendo utilizados, por exemplo, no Corpus de Referência do Português Contemporâneo (CRPC). Estes resultados foram subsequentemente alargados através de um outro projecto, o SemanticShare, em que se deu início à construção de um treebank, ou seja, à anotação do corpus com a representação sintática das frases.

Lançado em 2000, o Corpus de Extratos de Textos Eletrónicos MCT/Público (CETEMPúblico) é, por

sua vez, um corpus com cerca de 180 milhões de palavras provenientes de textos de um jornal diário português. A criação deste corpus teve como objetivo dar apoio ao desenvolvimento de ferramentas de processamento do português que necessitam de textos "em bruto" (i. e. sem anotação linguística) para a sua construção e avaliação. Este corpus foi criado no âmbito do projeto Processamento Computacional do Português, ao abrigo de um protocolo entre o Ministério da Ciência, Tecnologia e Ensino Superior e o jornal Público. Posteriormente, este projeto evoluiu para a Linguateca [49], um projeto de longo prazo para a tecnologia da linguagem do português.

Também em 2000, a tradução automática viria a ser o foco de um outro projecto apoiado pela Comissão Europeia, o TRADAUT, dirigido pela Universidade Nova de Lisboa. Este projecto teve por objectivo a melhoria da aplicação de tradução automática usada pelos serviços da Comissão Europeia para os pares de tradução entre o português, por um lado, e o inglês e o francês, por outro.

No campo do processamento de fala, cabe destaque para o projeto TECNOVOZ, iniciado em 2006. Este projeto foi liderado pelo INESC-ID e teve como objetivo principal favorecer a transferência de tecnologia para o setor empresarial, contando entre os seus parceiros com empresas como a estação de televisão pública RTP, entre outros. No setor empresarial, importa destacar a presença em Portugal, desde 2005, do Microsoft Language Development Center (MLDC), que tem igualmente contribuído para o desenvolvimento da indústria da tecnologia da linguagem no país.

Mais recentemente, instituições portuguesas e brasileiras têm participado no projeto CLARIN, que tem como objetivo a criação de uma infraestrutura de investigação europeia para a linguagem natural.

No Brasil, têm sido igualmente realizados esforços significativos em termos de investigação sobre tecnologia da linguagem para o português.

Como exemplos, pode referir-se a criação do Banco de Português no âmbito do projeto DIRECT, no início dos anos 90, pela Pontifícia Universidade Católica de São Paulo. Desde a sua criação, o Banco de Português tem sido uma importante fonte de dados para diversos estudos baseados em corpora.

Vale a pena referir também o corpus Summ-it, construído para dar apoio a estudos sobre sumarização automática, fenómenos anafóricos e relações retóricas no português. Este recurso foi desenvolvido no âmbito do projeto PLN-BR, do Núcleo Interinstitucional de Lingüística Computacional (NILC), levado a cabo pela Universidade de São Paulo e por um conjunto de investigadores de outras sete instituições brasileiras, em que foram produzidos uma série de outros corpora.

Mais recentemente, no período de 2006-2010, foi levado a efeito o projeto FAROL, liderado pela Universidade Pontifícia Católica do Rio Grande do Sul, que integrava quatro equipas de investigação. O objetivo principal deste projeto foi o reforço das ligações entre as diversas equipas, promovendo o intercâmbio entre estudantes e investigadores, de forma a melhorar a qualidade da investigação na área do processamento da linguagem natural.

A par de programas e projetos de investigação quer no Brasil quer em Portugal, cabe destacar o PROPOR enquanto principal iniciativa aglutinadora de uma crescente comunidade internacional de investigadores que trabalha sobre o português. O PROPOR é a conferência científica internacional de referência para o processamento computacional da língua portuguesa. É uma conferência bienal que desde 1993 tem lugar alternadamente nos dois países.

Estes são apenas alguns exemplos de iniciativas, projetos e programas na área da tecnologia da linguagem para a língua portuguesa. Representam um avanço importante. Existe ainda, porém, uma grande distância no que respeita à muito maior atividade de investigação sobre

outras línguas mais estudadas e para as quais o desenvolvimento de recursos linguísticos e tecnológicos se encontra muito mais avançado.

Comparado com o nível de financiamento para a tecnologia da linguagem não só para o inglês, mas também para idiomas até de bastante menor projeção global que a língua portuguesa, o apoio para a tecnologia da linguagem para o português é ainda muito baixo.

Em Portugal, o financiamento vem sobretudo do Ministério da Ciência, Tecnologia e Ensino Superior, através da Fundação para a Ciência e a Tecnologia (FCT). No entanto, a obtenção de apoios para projetos em tecnologia da linguagem tornou-se particularmente difícil, se não mesmo impossível, uma vez que as propostas de projetos nesta área são submetidas e avaliadas não na seção de Informática ou na de Ciências da Linguagem, mas na seção de Engenharia Eletrotécnica, em que têm de competir com centenas de propostas de projetos sobre assuntos completamente ortogonais e enfrentar um júri desconectado da área e dos seu temas.

Além da FCT, a Fundação Calouste Gulbenkian também financia, ocasionalmente, projetos na área da tecnologia da linguagem.

Comparado com o nível de financiamento para a tecnologia da linguagem não só para o inglês, mas inclusive para idiomas de bastante menor projeção global que a língua portuguesa, o apoio para a investigação sobre o português é ainda muito baixo.

No Brasil, embora ainda seja limitado, o financiamento para a investigação em geral, e para as atividades em tecnologia da linguagem em particular, vem sobretudo de agências governamentais. O Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a Financiadora de Estudos e Projetos (FINEP) são as quatro principais instituições de financiamento no país.

Algumas destas agências participaram inclusivamente em programas de financiamento conjunto com algumas empresas. Por exemplo, a FAPESP e o Microsoft Research Center formaram recentemente uma parceria para o financiamento de projetos socialmente relevantes no Estado de São Paulo, que incluiu, entre outros, o PorSimples [50], um projeto na área da tecnologia da linguagem que tem como objetivo a simplificação de textos de português para auxiliar leitores pouco alfabetizados a compreender textos da internet.

4.6 DISPONIBILIDADE DE FERRAMENTAS E RECURSOS

Nesta seção, é apresentado um resumo do estado atual da tecnologia da linguagem para o português. A Figura 8 contém o resultado de uma apreciação levada a efeito por especialistas na área quanto ao estado de desenvolvimento de recursos linguísticos e ferramentas de processamento para a língua portuguesa, com base numa escala de 0 (muito baixo) a 6 (muito alto) e de acordo com os sete critérios que encabeçam as colunas da figura.

Estes resultados devem ser apreciados no seguinte enquadramento:

- Apesar de haver sub-áreas muito ativas neste campo, em termos de tecnologia da linguagem, o português é um idioma menos bem equipado sobretudo quando comparado com línguas de países com uma aposta muito forte nesta tecnologia, como por exemplo, o inglês, o alemão ou o neerlandês;
- Foram compilados dois grandes corpora de texto "em bruto" para o português, sendo que um é pouco representativo, uma vez que abrange apenas um tipo de texto (jornalístico), e o outro não está totalmente disponível, devido a restrições de direitos de autor;

	Quantidade	Disponibilidade	Qualidade	Cobertura	Maturidade	Sustentabilidade	Adaptabilidade
Tecnologia da Linguagem: Ferramentas de Processamento e Aplicações							
Reconhecimento da Fala	2	3	4	2	2	2	4
Síntese da Fala	3	3	4	4	4	3	4
Análise Gramatical	3	3	4	4	4.5	2.5	4.5
Análise Semântica	1.5	2	3	2	2.5	2.5	2.5
Geração de Linguagem	0	0	0	0	0	0	0
Tradução Automática	3	2	2	2	4	2	2
Recursos Linguísticos: Conjuntos de Dados e Bases de Conhecimento Linguístico							
Corpora Escritos	3	3	4	4.5	4	4.5	4.5
Corpora de Fala	4	2	4	4	4	3	3
Corpora Paralelos	2	4	2	2	2	3	3
Recursos Lexicais	3.5	3	4.5	3	4	3	3
Gramáticas	1	4	5	2	2	2	2

8: Estado de desenvolvimento da tecnologia da linguagem para o português

- Está disponível um corpus anotado de 1 milhão de palavras, juntamente com o respetivo etiquetador morfossintático e outras ferramentas de processamento de base morfológica. Para as variedades do português menos estudadas, têm estado a ser construídos corpora nos últimos anos, que precisam porém de receber mais atenção;

- Em relação à tecnologia da fala, há um conjunto de sistemas comerciais para as variedades europeia e brasileira do português (reconhecimento da fala, síntese da fala e gestão de diálogo), e embora as equipas em Portugal e no Brasil sejam dinâmicas nesta área, as ferramentas e os corpora anotados não se encontram disponíveis, estando em regra reservados para uso interno dos laboratórios;

- É necessário bastante mais trabalho no desenvolvimento de recursos lexicais de todo o tipo, incluindo a criação de ontologias e a expansão de léxicos e wordnets, actualmente de volume muito reduzido;

- Não existem ainda corpora anotados com informação sobre semântica lexical, o que origina um preocupante entrave à investigação sobre desambiguação de aceção de palavras em português, assim como ao desenvolvimento de ferramentas associadas;

- Enquanto alguns corpora têm anotação morfossintática e outros tipos de informação morfológica, os corpora com anotação sintática (treebanks) são mais raros e de tamanho muito reduzido. Com base nestes recursos, têm sido desenvolvidos alguns analisadores sintáticos, que precisam porém de ser apro-

fundados. É necessário por isso bastante mais trabalho na construção de treebanks e no desenvolvimento de ferramentas de análise sintática.

- Quanto mais conhecimento linguístico e semântico uma ferramenta tomar em consideração, mais lacunas existem (ver, por exemplo, recuperação de informação vs. semântica do texto): é preciso aplicar mais esforço de Investigação e Desenvolvimento no processamento linguístico profundo, incluindo a construção de gramáticas computacionais para o português;
- As ferramentas de análise do texto e do discurso são poucas e foram alvo até agora de um desenvolvimento apenas parcial;
- Situação similar ou pior se encontra no que diz respeito a outras ferramentas ou aplicações de mais alto nível, como por exemplo, os sistemas de sumarização ou de resposta a perguntas, entre várias outras;
- Os corpora paralelos para tradução automática que incluem o português são, sobretudo, os disponibilizados por iniciativas desenvolvidas pela UE e, consequentemente, são muito limitados quanto ao domínio a que dizem respeito (e. g. texto jurídico).

Estes resultados da avaliação do estado de desenvolvimento da tecnologia da linguagem para o português apontam claramente para a necessidade premente de concentrar mais esforços quer na criação de recursos linguísticos quer na investigação de ferramentas para o processamento computacional do português e desenvolvimento de aplicações da tecnologia da linguagem.

Há uma 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.

A falta de dados em muito maior volume e a grande complexidade dos sistemas de tecnologia da linguagem tornam igualmente indispensável a criação de novas infraestruturas de investigação que apoiem a partilha de dados e estimulem a cooperação na investigação.

4.7 COMPARAÇÃO ENTRE LÍNGUAS

O estado atual de desenvolvimento da tecnologia da linguagem varia de forma significativa em função da língua em consideração. Para se obter uma ideia da situação entre as diferentes línguas, esta seção apresenta uma avaliação que tomou como amostra duas áreas de aplicação – a tradução automática e o processamento da fala – e uma tecnologia de base – a análise de texto –, assim como recursos de base (conjuntos de dados, bases de conhecimento linguístico, etc) necessários para a criação de ferramentas e aplicações em tecnologia da linguagem.
A classificação foi levada a efeito usando a seguinte escala:

1. Apoio excelente
2. Apoio bom
3. Apoio médio
4. Apoio fragmentário
5. Pouco ou nenhum apoio

O nível de apoio em termos de tecnologia da linguagem, classificado com essa escala, foi determinado de acordo com os seguintes critérios:
Tradução automática: Qualidade da tecnologia de tradução automática existente; número de pares de línguas cobertos; cobertura de fenómenos linguísticos e de domínios textuais; qualidade e tamanho dos corpora paralelos existentes; quantidade e variedade das aplicações de tradução automática.
Análise do Texto: Qualidade e cobertura da tecnologia do texto existente (morfologia, sintaxe, semântica);

cobertura em termos de fenómenos linguísticos e de domínios; quantidade e variedade das aplicações existentes; qualidade e tamanho dos corpora anotados; qualidade e cobertura dos recursos lexicais e das gramáticas existentes.

Processamento de fala: Qualidade da tecnologia de reconhecimento de fala existente; qualidade da tecnologia de síntese de fala; cobertura em termos de domínios; número e tamanho dos corpora de fala; quantidade e variedade das aplicações baseadas em tecnologia da fala.

Recursos: Qualidade e tamanho dos corpora escritos, de fala e paralelos existentes; qualidade e cobertura dos recursos lexicais e gramáticas.

As Figuras 9 a 12 mostram que a língua portuguesa está em posições um pouco diferentes consoante as áreas de investigação.

Quando comparada com o espanhol ou o italiano, por exemplo, a língua portuguesa está bem posicionada no que respeita às ferramentas e recursos da fala. Contudo, quanto a tradução automática, análise do texto e recursos linguísticos, o português está longe de contar com a mesma cobertura que o inglês (líder em quase todas as áreas da tecnologia da linguagem) e outras línguas, como por exemplo, o neerlandês ou o alemão, etc. Cabe porém não perder de vista que, até para o inglês, há ainda muitas lacunas, sobretudo no que diz respeito às aplicações de mais alto nível.

No caso do processamento da fala, a tecnologia atualmente existente tem um nível de desempenho suficiente para ser integrada em várias aplicações industriais, como os sistemas de diálogo ou de ditado.

As componentes de análise de texto e recursos linguísticos, por sua vez, já abrangem um leque considerável de fenómenos linguísticos e fazem parte de muitas aplicações que envolvem principalmente processamento superficial da linguagem natural, como por exemplo, a correção ortográfica ou as aplicações de apoio ao autor.

No entanto, para a construção de aplicações mais sofisticadas, como a tradução automática, os sistemas de resposta a perguntas, a sumarização, etc, existe uma clara necessidade de bastantes mais recursos e ferramentas, em quantidade e qualidade, que cubram uma mais ampla gama de aspetos linguísticos e que permitam uma análise mais profunda dos textos.

Ao melhorar a qualidade e a cobertura destes recursos e tecnologias de base, estar-se-á a criar novas oportunidades para aperfeiçoar um vasto leque de áreas de aplicação avançadas, incluindo a tradução automática abrangente e de alta qualidade.

4.8 CONCLUSÕES

Os resultados reunidos nesta coleção de Livros Brancos mostram que existem enormes diferenças entre as línguas europeias quanto à tecnologia da linguagem. Embora algumas línguas e áreas de aplicação estejam equipadas com software e recursos linguísticos em quantidade e qualidade, para outras línguas e aplicações, encontram-se várias lacunas, que em alguns casos podem ser muito significativas. Muitas línguas não estão ainda equipadas com a tecnologia básica para a análise de texto nem com os recursos linguísticos essenciais para o desenvolvimento dessa tecnologia. Outras línguas terão essas ferramentas e recursos básicos, mas a implementação de níveis de processamento mais avançados ainda se encontra a alguma distância. Nesta medida, é preciso realizar um esforço em grande escala para se alcançar o objetivo ambicioso de se assegurar tecnologia da linguagem de alta qualidade para todas as línguas, com especial destaque para a tradução automática de muito maior fiabilidade.

No caso do português, o apoio da tecnologia da linguagem para esta língua tem vindo a melhorar gradualmente, mas é necessário garantir o incremento estratégico do esforço aplicado nesta área para se vir a alcançar um patamar de desenvolvimento sustentado. Há uma boa comunidade de centros de investigação, tanto

em Portugal como no Brasil, que cooperam ativamente entre si e que, de momento, têm capacidade instalada para fazer avançar a tecnologia da linguagem para a língua portuguesa.

São porém necessárias medidas imediatas para que se possam obter progressos importantes para o português e assegurar a sua posição como língua de comunicação internacional com projeção global.

São necessárias medidas imediatas para que se possam obter progressos importantes para a língua portuguesa e assegurar a sua posição como língua de comunicação internacional com projeção global.

Tem-se registado uma falta de continuidade no financiamento da Investigação e Desenvolvimento. Programas de curta duração tendem a alternar com períodos de financiamento escasso ou mesmo nulo. A par disso, verifica-se ainda a conveniência de uma melhor coordenação de programas de investigação entre países, da Europa e de outros continentes, ou de articulação desses programas com programas da Comissão Europeia.

Os resultados deste livro apontam no sentido de que a única via de progresso consiste em se realizar um esforço substancial para se criarem recursos linguísticos para o português que permitam, por sua vez, impulsionar e fomentar a investigação, a inovação e o desenvolvimento de ferramentas e aplicações da tecnologia da linguagem.

A necessidade de grandes volumes de dados e a extrema complexidade dos sistemas da tecnologia da linguagem tornam também cruciais o desenvolvimento de uma infraestrutura e de uma organização de investigação mais coerente, que fomentem uma maior cooperação e partilha de resultados.

Apoio excelente	Apoio bom	Apoio médio	Apoio fragmentário	Pouco/nenhum apoio
	Inglês	Francês Espanhol	Alemão Catalão Húngaro Italiano Neerlandês Polaco Romeno	Basco Búlgaro Checo Croata Dinamarquês Eslovaco Esloveno Estónio Finlandês Galego Grego Irlandês Islandês Letão Lituano Maltês Norueguês **Português** Sérvio Sueco

9: Tradução Automática: estado da tecnologia da linguagem para 30 línguas europeias

Apoio excelente	Apoio bom	Apoio médio	Apoio fragmentário	Pouco/nenhum apoio
	Inglês	Alemão Espanhol Francês Italiano Neerlandês	Basco Búlgaro Catalão Checo Dinamarquês Eslovaco Esloveno Finlandês Galego Grego Húngaro Norueguês Polaco **Português** Romeno Sueco	Croata Estónio Irlandês Islandês Letão Lituano Maltês Sérvio

10: Análise do Texto: estado da tecnologia da linguagem para 30 línguas europeias

Apoio excelente	Apoio bom	Apoio médio	Apoio fragmentário	Pouco/nenhum apoio
	Inglês	Alemão Checo Espanhol Finlandês Francês Italiano Neerlandês **Português**	Basco Búlgaro Catalão Dinamarquês Eslovaco Esloveno Estónio Galego Grego Húngaro Irlandês Norueguês Polaco Sérvio Sueco	Croata Islandês Letão Lituano Maltês Romeno

11: Processamento da Fala: estado da tecnologia da linguagem para 30 línguas europeias

Apoio excelente	Apoio bom	Apoio médio	Apoio fragmentário	Pouco/Nenhum apoio
	Inglês	Alemão Checo Espanhol Francês Húngaro Italiano Neerlandês Polaco Sueco	Basco Búlgaro Catalão Croata Dinamarquês Eslovaco Esloveno Estónio Finlandês Galego Grego Norueguês **Português** Romeno Sérvio	Irlandês Islandês Letão Lituano Maltês

12: Recursos linguísticos escritos e orais: estado da tecnologia da linguagem para 30 línguas europeias

5

SOBRE A META-NET

A **META-NET** é uma Rede de Excelência para a investigação científica parcialmente financiada pela Comissão Europeia. A rede abrange atualmente 54 centros de investigação em 33 países da Europa. Resulta da agregação de quatro projetos europeus: CESAR, METANET4U, META-NORD e T4ME. O projeto METANET4U é coordenado pela Faculdade de Ciências da Universidade de Lisboa.

A META-NET promove a **META**, a Multilingual Europe Technology Alliance (Aliança Europeia para a Tecnologia Multilingue), uma comunidade com um número crescente de profissionais e de organizações da tecnologia da linguagem na Europa. A META-NET procura fazer avançar as fundações tecnológicas para uma sociedade europeia de informação verdadeiramente multilingue que:

- torne possíveis a comunicação e a cooperação usando-se línguas diferentes;
- assegure a todos os europeus o acesso à informação e ao conhecimento em igualdade de circunstâncias, independentemente da sua língua;
- desenvolva e melhore as funcionalidades da tecnologia de informação conetada em rede.

Esta Rede de Excelência contribui para o desenvolvimento de uma Europa que se une em torno de um espaço de informação digital único. Estimula e promove tecnologias multilingues para todas as línguas europeias. Estas tecnologias apoiam a tradução automática, a produção de conteúdos, o processamento de informação e a gestão do conhecimento para um amplo leque de domínios e aplicações. Tornam também possíveis interfaces intuitivas baseadas em linguagem que permitem a interação com os mais diversos dispositivos, que abrangem desde os eletrodomésticos até maquinaria e veículos, incluindo, entre vários outros, computadores e robôs.

Lançada a 1 de fevereiro de 2010, a META-NET já realizou várias atividades nas suas três linhas de ação: META-VISION, META-SHARE e META-RESEARCH.

A **META-VISION** promove uma comunidade dinâmica e influente de atores que se unem em torno de uma perspetiva partilhada e de uma Agenda de Investigação Estratégica (AIE) comum. O enfoque principal desta linha de ação consiste no desenvolvimento, na Europa, de uma comunidade coerente e coesa que se reúne em torno da tecnologia da linguagem, juntando representantes de grupos altamente fragmentados e diversificados de atores. O presente Livro Branco foi preparado juntamente com volumes similares para outras 29 línguas. A perspetiva partilhada acerca da tecnologia foi desenvolvida em três Grupos de Perspetiva setoriais. O META Technology Council foi constituído para discutir e preparar a AIE baseada nessa perspetiva partilhada, através de uma interação intensa com toda a comunidade da tecnologia da linguagem.

A **META-SHARE** cria uma plataforma, aberta e distribuída, para a permuta e partilha de recursos linguísticos. A rede peer-to-peer de repositórios conterá dados linguísticos, ferramentas e serviços web, que são documentados com metadados de elevada qualidade e organizados em categorias padronizadas. O recursos po-

dem ser acedidos de forma imediata e estão organizados de forma a permitir que sobre eles se efetuem pesquisas de maneira uniforme. Os recursos disponíveis incluem materiais gratuitos e de código aberto, assim como elementos restritos, de natureza comercial, que podem ser adquiridos.

A **META-RESEARCH** constrói pontes em direção a áreas tecnológicas relacionadas. Esta atividade procura estimular avanços noutros campos e tirar partido de investigação inovadora que possa beneficiar a tecnologia da linguagem. Em particular, esta linha de ação focase: na realização de investigação de ponta em tradução automática; na angariação de dados; na preparação de conjuntos de dados e organização de recursos linguísticos tendo em vista processos de avaliação; na compilação de inventários de ferramentas e métodos; e na organização de workshops e eventos de formação para membros da comunidade.