D2.3.pt
Language Report for Portuguese
(Portuguese version)

Version 1.0

2011-07-29

METANET4U

www.metanet4u.eu

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.
This central objective is articulated in terms of the following main goals:
Assessment: to collect, organize and disseminate information that permits an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programs; prevailing standards and practices; current level of development, main drivers and roadblocks; etc.
Collection: to assemble and prepare language resources for distribution. This includes collecting languages resources; documenting these language resources; upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.
Distribution: to distribute the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaboration with other projects and, where useful, with other relevant multi-national forums or activities. It also includes helping to build and operate broad inter-connected repositories and exchange facilities.
Dissemination: to mobilize national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

METANET4U is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META. META is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society.

METANET4U is co-funded by the participating institutions and the ICT Policy Support Programme of the European Commission

and by the participating institutions:

Faculty of Sciences, University of Lisbon

Instituto Superior Técnico

University of Manchester

University *Alexandru Ioan Cuza*

Research Institute for Artificial Intelligence, Romanian Academy

University of Malta

Technical University of Catalonia

Universitat Pompeu Fabra

Revision History

Version	Date	Author	Organisation	Description
0.1	18-07-2011	Amália Mendes, António Branco, Paulo Henriques, Sílvia Pereira, Isabel Trancoso, Thomas Pellegrini, Hugo Meinedo, Paulo Quaresma	ULX, IST	Draft version
0.2	28-07-2011	Núria Bel	UPF	Review notes
1.0	29-07-2011	Amália Mendes, António Branco, Paulo Henriques, Sílvia Pereira, Isabel Trancoso, Thomas Pellegrini, Hugo Meinedo, Paulo Quaresma	ULX, IST	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

METANET4U

D2.3.pt
Language Report for Portuguese
(Portuguese version)

Document METANET4U-2011-D2.3.pt
EC CIP project #270893

Deliverable D2.3.pt
Completion: Final
Status: Submitted
Dissemination level: Public

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: ULX, IST

Authors: Amália Mendes, António Branco, Paulo Henriques, Sílvia Pereira, Isabel Trancoso, Thomas Pellegrini, Hugo Meinedo, Paulo Quaresma

Reviewer: Núria Bel

META-NET Série Livro Branco

As Línguas na Sociedade Europeia de Informação

– Português –

A realização deste livro branco foi financiada pelo 7.º Programa-Quadro do ICT Policy Support Programme 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).

Este livro branco faz parte de uma série que promove o conhecimento sobre as tecnologias da linguagem e o seu potencial. É particularmente dirigido a educadores, jornalistas, políticos, comunidades linguísticas, entre outros.

A disponibilidade e a utilização das tecnologias da linguagem na Europa variam consoante as línguas. Por conseguinte, as acções que são necessárias para apoiar a investigação e o desenvolvimento destas tecnologias também são diferentes para cada língua. Estas acções dependem de vários factores, tais como a complexidade de uma determinada língua e a dimensão da sua comunidade.

META-NET, um projecto da Rede de Excelência da Comissão Europeia, levou a cabo uma análise sobre os actuais recursos linguísticos e tecnologias de 23 línguas oficiais europeias, bem como de outras línguas importantes a nível nacional e regional na Europa. Os resultados desta análise sugerem a existência de significativas lacunas de investigação para cada língua. Análises e avaliações mais detalhadas da situação actual poderão ajudar a maximizar o impacto de investigação adicional e minimizar os riscos.

O projecto META-NET abrange 47 centros de investigação de 31 países, que estão a trabalhar com as partes interessadas de empresas comerciais, agências governamentais, indústria, instituições de investigação, empresas de *software*, fornecedores de tecnologia e universidades europeias. Em conjunto, pretende-se criar uma visão tecnológica comum, desenvolvendo, ao mesmo tempo, uma agenda estratégica de investigação que mostrará de que modo as aplicações na área das tecnologias da linguagem poderão vir a suprir algumas lacunas na investigação até 2020.

META-NET
DFKI Projektbüro Berlin
Alt-Moabit 91c
10559 Berlin
Germany

office@meta-net.eu
http://www.meta-net.eu

Autores

Doutora Amália Mendes, CLUL-University of Lisbon
Prof. Doutor António Branco, University of Lisbon
Dr. Paulo Henriques, CLUL-University of Lisbon
Dr.ª Sílvia Pereira, University of Lisbon
Prof.ª Doutora Isabel Trancoso, INESC-ID / IST
Doutor Thomas Pellegrini, INESC-ID
Doutor Hugo Meinedo, INESC-ID
Prof. Doutor Paulo Quaresma, University of Évora

Agradecimentos

Gostaríamos de agradecer aos autores do livro branco alemão por terem autorizado a reprodução de materiais da sua publicação.

Índice

Sumário

Muitas das línguas europeias correm actualmente o risco de se tornarem vítimas da era digital por estarem pouco representadas na internet e aí disponibilizarem poucos recursos. Existem imensas oportunidades nos mercados regionais que são pouco exploradas devido às barreiras linguísticas. Se não forem tomadas medidas, em breve muitos cidadãos europeus serão social e economicamente prejudicados por falarem a sua língua materna.

As Tecnologias da Linguagem (TL) são um meio que permitirá aos cidadãos europeus a participação numa sociedade da informação e do conhecimento inclusiva, igualitária e economicamente bem sucedida. As Tecnologias da Linguagem multilingues serão um portal para uma comunicação e interacção instantânea, menos dispendiosa e sem esforço entre falantes de línguas diferentes.

Actualmente, os serviços na área das Tecnologias da Linguagem são sobretudo oferecidos por fornecedores comerciais com sede nos Estados Unidos, como por exemplo o serviço gratuito de tradução do Google. Outro exemplo, que mostra ainda as enormes potencialidades das Tecnologias da Linguagem, é o super computador Watson da IBM, que recentemente venceu um episódio do jogo Jeopardy, enfrentando concorrentes humanos.

Enquanto europeus, temos de nos colocar algumas questões importantes:

- Devem as nossas comunicações e infraestruturas do conhecimento estar dependentes de empresas monopolistas?
- Podemos, de facto, confiar em serviços na área da linguagem cuja disponibilização depende inteiramente de terceiros?
- Estamos realmente a competir no mercado global para a investigação e para o desenvolvimento das Tecnologias da Linguagem?
- Estarão outras instituições de outros continentes dispostas a abordar e a tratar os nossos problemas de tradução e outras questões relacionadas com o multilinguismo europeu?
- Poderá o nosso contexto cultural europeu ajudar a construir a sociedade do conhecimento, ao contribuir com tecnologias de alta qualidade, mais seguras, precisas, inovadoras e robustas?

Neste livro branco para a língua portuguesa fica demonstrado que na área da investigação em Portugal existe um ambiente dinâmico, que precisa de ser estimulado para oferecer um bom suporte a uma indústria emergente na área das Tecnologias da Linguagem. Apesar de um conjunto importante de recursos linguísticos e ferramentas de processamento ter sido desenvolvido para o Português, existe ainda um menor número de soluções se compararmos com outras línguas da União Europeia.

De acordo com a avaliação detalhada apresentada neste Relatório, devem ser tomadas medidas imediatas para que possam ser alcançados progressos significativos para a língua portuguesa.

Um risco para as nossas línguas e um desafio para as Tecnologias da Linguagem

Estamos neste momento a testemunhar uma revolução digital, comparável à invenção da imprensa por Gutenberg.

Somos hoje testemunhas de uma revolução digital que está a ter um impacto radical na comunicação e na sociedade. Os recentes desenvolvimentos nas tecnologias da comunicação digitais em rede 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?

Depois da invenção de Gutenberg, os avanços na comunicação e na partilha de conhecimentos foram realizados mediante alguns esforços, como a tradução da Bíblia do Latim para novas línguas europeias, por Lutero. 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 perspectivas 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 permitiu uma partilha de conhecimento entre línguas;
- a criação de directrizes jornalísticas 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 formatos, veio dar resposta às diferentes necessidades de comunicação.

Nos últimos vinte anos, as tecnologias da informação ajudaram a automatizar e a facilitar muitos dos processos:

- Os programas avançados de edição de texto (*desktop publishing software*) substituem a dactilografia e digitação;
- as projecções de transparências dão lugar aos powerpoints;
- o correio electrónico permite receber e enviar documentos de forma mais rápida do que o fax;
- o Skype é utilizado para a realização de chamadas de telefone pela internet e para videoconferências;
- os formatos de codificação de áudio e vídeo facilitam a troca de conteúdos multimédia;
- os motores de busca sugerem palavras-chave baseadas nos acessos a páginas Web;
- os serviços online, como o serviço gratuito de tradução do Google, produzem traduções de uma forma rápida e aproximada;
- as plataformas de redes sociais facilitam a colaboração e a partilha de informação.

Apesar de estas ferramentas e aplicações serem úteis, elas não são actualmente suficientes para implementar uma sociedade de informação europeia multilingue e sustentável, uma sociedade moderna e inclusiva onde informação e bens possam fluir livremente.

Fronteiras linguísticas travam a sociedade de informação europeia

Não podemos saber exactamente como será o futuro da sociedade de informação. Por exemplo, quando se pretende discutir uma estratégia europeia comum em política externa, talvez queiramos ouvir os Ministros dos Negócios Estrangeiros falar na sua língua materna. Poderemos querer uma plataforma onde pessoas que falam diferentes línguas e têm diferentes competências linguísticas possam discutir um determinado assunto enquanto automaticamente a tecnologia reúne as suas opiniões e produz breves resumos. Também podemos ter de falar com a linha de apoio de um seguro de saúde localizado num país estrangeiro.

As necessidades de comunicação são de facto muito diferentes quando comparadas com o que acontecia alguns anos atrás. Numa economia global e num espaço de informação com mais línguas, falantes e conteúdos, somos levados a interagir mais rapidamente com novos tipos de meios de comunicação. A recente popularidade das redes sociais (Wikipédia, Facebook, Twitter e YouTube) é apenas a ponta do iceberg.

A nova economia global e a nova sociedade de informação confrontam-nos com mais línguas, mais falantes e mais conteúdos.

Hoje podemos transmitir gigabytes de texto para todo o mundo em poucos segundos antes de percebermos que está numa língua que não entendemos. De acordo com um recente relatório solicitado pela 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, seguida pelo francês, alemão e espanhol.) 55% dos utilizadores lêem conteúdos numa língua estrangeira, enquanto apenas 35% utilizam outra língua para escrever emails ou colocar comentários na internet[i]. Há alguns anos atrás o inglês pode ter sido 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 explodiu (em particular línguas asiáticas e árabes).

É surpreendente que esta divisão digital provocada por fronteiras linguísticas não receba muita atenção nos discursos políticos; ainda assim levanta uma questão premente: "Que línguas europeias se vão manter e prosperar na sociedade do conhecimento e da informação em rede?"

Que línguas europeias se vão manter e prosperar na na sociedade do conhecimento e da informação em rede?

As nossas línguas em risco

Na Europa, a imprensa escrita contribuiu para uma valiosa partilha de informação, mas também levou à extinção de muitas línguas europeias. Línguas regionais e minoritárias raramente foram impressas. Como consequência, muitas línguas, como o *Cornish* ou o Dálmata, foram muitas vezes reduzidas a formas orais de transmissão, o que limitou a sua adopção, disseminação e uso.

As cerca de 60 línguas da Europa são um dos mais ricos e importantes bens culturais que possuimos. A multiplicidade de línguas é também uma parte vital do êxito social da Europa[ii]. Enquanto línguas populares, como o inglês e o espanhol, manterão a sua presença na sociedade e no mercado digitais emergentes, muitas línguas europeias poderão desaparecer da comunicação digital e tornar-se irrelevantes para a sociedade de informação. Tais desenvolvimentos seriam certamente indesejáveis. Por um lado, estaria perdida uma oportunidade estratégica que enfraqueceria a posição global da Europa. Por outro lado, essa evolução entraria em conflito com o objectivo da participação igualitária de todos os cidadãos europeus independentemente da

A multiplicidade de línguas é também uma parte vital do êxito social da Europa.

língua. 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 em sociedade.[iii]

A tecnologia da linguagem é uma tecnologia-chave

No passado, os esforços de investimento concentraram-se no ensino das línguas e na tradução. Por exemplo, de acordo com algumas estimativas, o Mercado Europeu de tradução, interpretação, localização de *software* e preparação de websites para o mercado global (*website globalisation*) foi de 8,4 biliões de euros em 2008 e deverá crescer 10% por ano.[iv] Contudo, esta capacidade de crescimento não é suficiente para satisfazer as necessidades actuais e futuras.

As Tecnologias da Linguagem são uma tecnologia-chave que permite proteger e promover as línguas europeias. As Tecnologias da Linguagem ajudam as pessoas a colaborar, a conduzir negócios, a partilhar conhecimentos e a participar em debates sociais e políticos, independentemente das barreiras linguísticas ou das competências informáticas. As Tecnologias da Linguagem já auxiliam as tarefas do dia-a-dia, como escrever mensagens de correio electrónico, realizar pesquisas online ou reservar voos. Estamos a beneficiar das Tecnologias da Linguagem quando:

As Tecnologias da Linguagem ajudam as pessoas a colaborar, a conduzir negócios, a partilhar conhecimentos e a participar em debates sociais e políticos, em diferentes línguas.

- encontramos informação na internet, através de um motor de busca;
- verificamos a ortografia e a gramática num processador de texto;
- vemos as recomendações para um produto numa loja *online*;
- ouvimos as indicações verbais de um sistema de navegação;
- traduzimos páginas da internet com um serviço *online*.

As Tecnologias da Linguagem descritas em pormenor neste relatório são parte essencial de aplicações inovadoras no futuro. As Tecnologias da Linguagem são tipicamente uma tecnologia que permite aplicações num quadro mais vasto, como um sistema de navegação ou um motor de busca. Estes livros brancos focam-se na disponibilidade de tecnologias de base para cada língua.

Num futuro próximo, precisaremos, para todas as línguas europeias, de Tecnologias da Linguagem disponíveis, acessíveis e totalmente integradas em ambientes informáticos mais latos. Sem elas, os utilizadores não poderão ter acesso a experiências de tipo interactivo, multimédia e multilingue.

Oportunidades para as Tecnologias da Linguagem

As Tecnologias da Linguagem permitem fazer tradução automática de conteúdos, produção, processamento, informação e gestão do conhecimento para todas as línguas europeias. Também podem ajudar no desenvolvimento de interfaces intuitivos, dirigidos a cada língua específica, para electrodomésticos, máquinas, veículos, computadores e robôs. Embora muitos protótipos já existam, as aplicações comerciais e industriais ainda estão nas primeiras fases de desenvolvimento. Resultados recentes em investigação e desenvolvimento criaram uma verdadeira janela de oportunidades. Por exemplo, a Tradução Automática (TA) já oferece um nível razoável de precisão dentro de domínios específicos, e certas

aplicações experimentais fornecem informação, gestão do conhecimento e produção de conteúdos em muitas línguas europeias.

As aplicações no âmbito da linguagem, as interfaces baseadas na voz do utilizador e os sistemas de diálogo são habitualmente encontrados em domínios altamente especializados, e muitas vezes apresentam um desempenho limitado. Por exemplo, um campo activo da investigação é o uso das Tecnologias da Linguagem para as operações de salvamento em áreas de desastres. Nestes ambientes de alto risco, a precisão da tradução pode significar a diferença entre a vida e a morte. O mesmo raciocínio se aplica à utilização das Tecnologias da Linguagem na indústria dos cuidados de saúde. Robôs inteligentes com capacidades em várias línguas têm o potencial de salvar vidas.

No mercado das indústrias da educação e do entretenimento, existem grandes oportunidades para a integração das Tecnologias da Linguagem em jogos, em produtos que associam educação e entretenimento, em ambientes de simulação ou programas de treino. Os serviços de informação móvel, os programas de aprendizagem de uma língua assistida por computador, os ambientes de *e-learning*, as ferramentas de auto-avaliação e os programas de detecção de plágios, são apenas alguns dos exemplos onde estas tecnologias podem desempenhar um papel importante. A popularidade das redes sociais, como o Twitter e o Facebook, sugerem uma maior necessidade de sofisticação das Tecnologias da Linguagem para permitir uma monitorização de mensagens, resumir discussões, sugerir tendências de opinião, detectar respostas emocionais, identificar infracções aos direitos de autor ou seguir pistas de uso indevido.

Multilingualismo é a regra, não a excepção.

As Tecnologias da Linguagem representam uma enorme oportunidade para a União Europeia, tanto a nível económico como cultural. O multilinguismo na Europa tornou-se regra. As empresas europeias, organizações e escolas também se tornaram multinacionais e diversificadas. Os cidadãos querem comunicar entre si, para além das fronteiras linguísticas que ainda existem no Mercado Comum Europeu. As Tecnologias da Linguagem podem ajudar a ultrapassar os obstáculos que restam, permitindo o uso livre e aberto da língua. Além disso, Tecnologias de Linguagem multilingues e inovadoras para a Europa podem ajudar-nos a comunicar com os nossos parceiros globais e as suas comunidades multilingues. As Tecnologias da Linguagem representam uma riqueza de oportunidades económicas internacionais.

Desafios para as Tecnologias da Linguagem

O actual ritmo de progresso tecnológico e inovação de produtos é muito lento e não podemos esperar 10 ou 20 anos para haver melhorias significativas.

Embora as Tecnologias da Linguagem tenham feito progressos consideráveis nos últimos anos, o actual ritmo de progresso tecnológico e de inovação de produtos é muito lento. Não podemos esperar 10 ou 20 anos para obter melhorias significativas que possam promover a comunicação e a produtividade no nosso ambiente multilingue.

As Tecnologias da Linguagem com maior utilização, como a verificação da gramática e ortografia em processadores de texto, são normalmente monolingues e estão apenas disponíveis para um pequeno conjunto de idiomas. As aplicações para uma comunicação multilingue requerem um certo nível de sofisticação. Os tradutores automáticos e serviços online, como o Google Translate ou o Bing Translator, são excelentes na criação de aproximações ao conteúdo dos documentos. Mas estes serviços

online e estas aplicações de tradução automática profissionais apresentam várias dificuldades quando se exigem traduções mais complexas e precisas. Existem muitos exemplos bem conhecidos e caricatos de erros de tradução, por exemplo a tradução literal dos nomes *Bush* e *Kohl*, que mostram os desafios que as Tecnologias da Linguagem ainda têm que enfrentar.

Aquisição da linguagem

Para ilustrar como os computadores lidam com a linguagem e como a apreensão da linguagem é uma tarefa difícil, fazemos uma breve abordagem sobre a forma como os seres humanos adquirem as suas primeira e segunda línguas; seguidamente esboçamos a forma como os sistemas de tradução automática funcionam – note-se que há uma razão para o facto de o domínio da tecnologia da linguagem estar intimamente ligado ao campo da inteligência artificial.

Os seres humanos adquirem competências linguísticas de dois modos diferentes. Primeiro, um bebé aprende uma língua ouvindo a interacção entre os falantes dessa língua. A exposição a exemplos linguísticos por falantes da língua, como pais, irmãos e outros membros da família, ajuda os bebés a partir dos dois ou mais anos de idade a produzir as suas primeiras palavras e frases curtas. Isso só é possível devido à predisposição genética dos humanos para a aprendizagem de línguas.

Os seres humanos adquirem competências linguísticas de dois modos diferentes: através de exemplos e através de regras linguísticas

Aprender uma segunda língua geralmente requer um esforço maior quando uma criança não está inserida numa comunidade de falantes dessa língua. Em idade escolar, as línguas estrangeiras são normalmente aprendidas através de descrições da sua estrutura gramatical, vocabulário e ortografia a partir de livros e materiais didácticos que descrevem o conhecimento linguístico em termos de regras abstractas, tabelas e exemplos. Aprender uma língua estrangeira exige muito tempo e esforço e torna-se mais difícil com a idade.

Os dois principais sistemas de Tecnologias da Linguagem adquirem capacidades linguísticas de forma semelhante à dos humanos. As abordagens estatísticas obtêm conhecimentos linguísticos a partir de vastas colecções de exemplos concretos de textos num único idioma ou em textos paralelos, isto é, textos disponíveis em duas ou mais línguas. Máquinas de aprendizagem de algoritmos modelam uma espécie de faculdade da linguagem capaz de derivar padrões de como as palavras, frases curtas e frases completas são corretamente utilizadas num único idioma ou correctamente traduzidos de uma língua para outra. As abordagens estatísticas exigem um elevado número de frases e a qualidade do seu desempenho aumenta à medida que aumenta o número de textos analisados. Não é incomum treinar tais sistemas a partir de dados que contêm milhões de frases. Esta é uma das razões pelas quais os fornecedores de motores de busca estão ansiosos por recolher o máximo de material escrito possível. A correcção ortográfica em processadores de texto, as informações disponíveis on-line e serviços como o Google Search e Google Translate dependem de uma abordagem estatística, isto é, orientada para os dados.

Os dois principais sistemas de Tecnologias da Linguagem adquirem capacidades linguísticas de forma semelhante à dos humanos.

Os sistemas baseados em regras são o segundo maior tipo de tecnologia da linguagem. Especialistas em Linguística, Linguística Computacional e Ciências Computacionais codificam a análise gramatical e compilam listas de vocabulário (léxicos). O desenvolvimento de um sistema baseado em regras é muito exigente em termos de tempo e trabalho e requer um conjunto de

especialistas. Alguns dos principais sistemas baseados em regras de tradução automática estão em constante desenvolvimento há mais de 20 anos. A vantagem de sistemas baseados em regras é que os peritos têm um controlo mais detalhado sobre o processamento da linguagem. Isto torna possível corrigir de forma sistemática os erros no programa e dar uma resposta detalhada ao utilizador, especialmente quando sistemas baseados em regras são usados para a aprendizagem de línguas. Devido a limitações financeiras, as Tecnologias da linguagem baseadas em regras só são viáveis para os principais idiomas.

O português na sociedade de informação europeia

Factos gerais

O português é a terceira língua europeia no mundo, com cerca de 200 milhões de falantes nativos, e um total de 220 milhões de falantes (como língua materna e segunda língua) em 4 continentes: Europa, América, África e Ásia[v].

É a língua oficial de Portugal, Brasil, Angola, Cabo Verde, Guiné-Bissau, Moçambique, São Tomé e Príncipe, Timor-Lorosae, Macau, Goa e, desde 2010, da Guiné Equatorial.

Devido ao fluxo de emigração [vi], o português é também falado pelas comunidades portuguesas em muitos países, ocupando em alguns deles uma importante posição entre o ranking da 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[vii].

O português é uma das línguas oficiais da União Europeia, do Mercosul (União económica que inclui o Brasil e outros países sul-americanos) e da União Africana. Com o desenvolvimento da alfabetização nas ex-colónias em África e Timor-Leste, o português tem um grande potencial de crescimento como segunda língua.

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

A divisão geográfica dos dialectos em Portugal[ix] distingue os dialectos do Centro-Sul dos dialectos do Norte. As diferenças são facilmente identificáveis ao nível da fonética e fonologia e ao nível lexical, embora todos os dialectos sejam mutuamente compreensíveis (possivelmente com a excepção de alguns dialectos das ilhas dos Açores). Os dialectos do Norte podem ser distinguidos pela ausência da distinção fonológica entre /b/ e /v/, com prevalência do /b/; pela preservação de antigos ditongos; e pela existência de fricativas ápico-alveolares.

Dada a dimensão geográfica do Brasil, não é viável apresentar aqui as suas variedades linguísticas. Por razões geográficas, políticas e sociais, nem é possível falar de uma variante padrão do português do Brasil. Em vez disso, os especialistas tendem a falar em "normas urbanas cultas", que apresentam algumas diferenças relativamente à norma do português europeu.

As variedades africanas do português também são diferentes da europeia, mas em menor grau, e partilham algumas características com o português do Brasil. A situação das variedades africanas diverge muito entre si: enquanto em Angola e Moçambique o número de falantes de português tem vindo a aumentar desde a independência em relação a Portugal, noutros casos, como São Tomé e Princípe e Cabo Verde, utiliza-se amplamente o crioulo, e o português é a segunda língua.

Particularidades da língua portuguesa

O português é uma língua românica, pelo que a maioria do seu léxico deriva do Latim. Adoptou também muitas palavras de uma grande variedade de outras línguas, em diferentes momentos, que, em muitos casos, permanecem entre as palavras mais frequentes (exemplos pré-latinos: *barranco, seara, bruxa*; germânicos: *luvas, bando, guerra*; e principalmente árabe: *aldeia, açúcar, laranja*).

A língua portuguesa pode muitas vezes soar para um ouvinte estrangeiro como uma sequência de consoantes. Isto deve-se ao facto de, diferentemente das outras línguas românicas, as vogais átonas do português serem muitas vezes enfraquecidas ou mesmo não pronunciadas. Este enfraquecimento das vogais é uma mudança tardia no português europeu e não afecta a variedade falada no Brasil.

A ordem básica das palavras em português é SVO – Sujeito Verbo Objecto (*Ele leu o livro*). Em alguns contextos pragmáticos (por exemplo, leitura enfática), a ordem VSO também ocorre (*lês tu o livro*) e as ordens OSV ou OVS são possíveis em construções marcadas como frases topicalizadas (*O livro, ele não leu*).

O português é uma língua de sujeito nulo, isto é, o sujeito da frase pode não estar realizado foneticamente (*Ø li o livro*). Quando o sujeito é um pronome de primeira pessoa, a sua não-realização é, de facto, a opção por defeito e habitualmente não há nenhum pronome expletivo nas construções impessoais (*Ø há um livro sobre esse tema*). Esta característica do português representa um desafio específico para a análise sintáctica automática dos textos e do discurso do português.

O paradigma flexional do português é muito mais rico do que o do inglês, particularmente no caso dos verbos: por exemplo, um verbo que segue um paradigma regular terá diferentes marcas para aspecto/ tempo/ modo, pessoa e número, atingindo mais de 70 formas flexionadas diferentes.

Além disso, há dois paradigmas de flexão verbal que não existem nas outras línguas românicas oficiais e 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 os marcadores de aspecto/tempo/constituinte modal e pessoa/número (*para eu cantar, para tu cantares, para eles cantarem*). As formas flexionadas do futuro do conjuntivo são homónimas com as do infinitivo não flexionado, excepto com os verbos irregulares, e isto aumenta o número de formas ambíguas no paradigma do verbo.

A posição dos pronomes clíticos na frase é outra característica que põe desafios específicos ao processamento automático da língua portuguesa. Os pronomes clíticos podem ocorrer antes e depois do verbo, mas nos tempos futuro e condicional, também podem aparecer no meio da forma verbal (*cantar-lhe-ei uma canção*). Além disso, a presença de um clítico de terceira pessoa no meio da forma verbal (e também no final) pode afectar o verbo: por exemplo, na sequência final -ar, o -r cai e a vogal é acentuada (*cantá-la-ei*).

Desenvolvimentos recentes

Sendo o inglês a língua mais difundida no mundo, a sua influência noutras línguas, incluindo o português, é cada vez mais perceptível. O cinema e a televisão, sobretudo séries norte-americanas, a música e a internet, acabam por abrir uma janela à presença

regular da língua inglesa no quotidiano e muitas palavras acabam por ser integradas na língua portuguesa.

É sobretudo na linguagem dos negócios e na internet que as palavras inglesas ganham visibilidade, como *CEO, stock, manager, briefing, casual day* ou *download, pen USB, delete, upload, refresh, online, site* e também *lifting, e-learning, shopping*. A influência do inglês sente-se tanto no português europeu como em outras variedades do português no mundo, nomeadamente na brasileira, que tende a adaptar estas palavras por empréstimo, como *deletar, googlar* e *twitar*.

No que diz respeito à música, embora haja muitos projectos musicais com letras em inglês dirigidos a um público mais jovem, os projectos cantados em português como o *Fado* e outros tipos de música tradicional portuguesa, e que foram considerados durante algum tempo fora de moda pelo público mais jovem, estão agora a recuperar uma grande audiência de todas as idades, tendo fortes reflexos na língua portuguesa.

Na última década tem havido um crescimento da relevância do português no contexto económico internacional, sobretudo devido ao desenvolvimento económico do Brasil. 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 reflecte-se no número crescente de pessoas que se inscrevem em cursos de português por todo o mundo.

O divulgação da língua em Portugal e no estrangeiro

Não há nenhuma instituição com a função de estabelecer a norma para a língua portuguesa, ao contrário do francês (Académie Française) e do espanhol (Real Academia Española), por exemplo. A Academia das Ciências de Lisboa e a Academia Brasileira das Letras dão algum contributo, em particular com a publicação de dicionários de referência: o Dicionário da Língua Portuguesa Contemporânea, em Portugal, e o Dicionário da Academia Brasileira de Letras, no Brasil.

O Instituto Camões é uma instituição sob a tutela do Ministério dos Negócios Estrangeiros e o seu principal objectivo é a promoção do português no mundo, fornecendo apoio a actividades culturais relacionadas com a língua, através da concessão de bolsas de estudo a nacionais e estrangeiros e apoiando o português como língua de comunicação internacional, particularmente em instituições internacionais como as Nações Unidas. Esta instituição também coordena e apoia o ensino do português em universidades e centros de cultura e língua portuguesa no estrangeiro.

A Comunidade dos Países de Língua Oficial Portuguesa (CPLP) é uma organização inter-governamental para a cooperação e que tem tido um papel activo na divulgação e promoção da língua portuguesa. O Instituto Internacional da Língua Portuguesa foi criado no âmbito da CPLP, mas ainda aguarda um maior empenho por parte dos decisores políticos. Foi também no seio da CPLP que foram empreendidos esforços para a preparação de um Novo Acordo Ortográfico, de forma a apoiar a expansão da língua e a sua consolidação no cenário económico e político internacional. Depois de alguma resistência inicial, este Novo Acordo Ortográfico

(iniciado em 1990) inclui todos os países que têm o português como língua oficial: Portugal, Brasil, Angola, Moçambique, Guiné-Bissau, Cabo Verde, São Tomé e Princípe e Timor-Leste.

A rádio e televisão públicas de Portugal têm-se empenhado na promoção do português através da transmissão de programas que procuram ensinar as boas práticas do uso do português padrão. Por exemplo, o programa semanal "Cuidado com a Língua" é simultaneamente educativo e divertido e divulga o Novo Acordo Ortográfico. Também a televisão e rádio públicas emitem diariamente um curto programa para esclarecer algumas dúvidas frequentes sobre a norma do português, e na rádio pública há debates regulares sobre as boas práticas do português escrito e falado. Há também muitas publicações dedicadas à salvaguarda da língua portuguesa, procurando atrair mais público para um uso adequado desta língua. Todos estes programas e publicações têm despertado o interesse da população pelas questões da língua.

No sector da música o uso do português tem sido apoiado através de um sistema de quotas nas rádios portuguesas, sendo que a lei estipula uma percentagem obrigatória de música em português nos programas emitidos. Inicialmente esta lei tinha estipulado uma quota de 25% a 40% e acabou por se fixar nos 25%.

A língua portuguesa também é promovida através do aumento da projecção internacional de autores portugueses, como os filósofos José Gil e Eduardo Lourenço, assim como os escritores António Lobo Antunes, Gonçalo M. Tavares, José Luís Peixoto, e o recentemente desaparecido Prémio Nobel José Saramago, cujas obras foram traduzidas em todo o mundo.

A língua e a educação

Nos últimos anos houve um grande investimento no desenvolvimento de uma rede de bibliotecas escolares. Isto foi feito no âmbito do Plano Nacional de Leitura, cujo objectivo-chave é a promoção dos índices de literacia dos estudantes portugueses nos vários níveis de aprendizagem, mas com enfoque principal nos primeiros anos de ensino.

Outra iniciativa recente foi a integração generalizada das novas Tecnologias da Informação nas escolas. Os alunos mais novos têm a possibilidade de adquirir, a baixo custo, computadores portáteis especialmente concebidos para os diferentes níveis de ensino. Em conjunto com este acesso a computadores pessoais, foram desenvolvidos softwares de programas educacionais, em que o português é a língua utilizada, e em que, em muitos casos, a aprendizagem da gramática é particularmente incentivada. Os resultados alcançados pelos alunos nos próximos anos permitirão uma avaliação aprofundada deste grande investimento em novas tecnologias.

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.

Será importante seguir num futuro próximo o impacto deste investimento no Plano Nacional de Leitura e nas novas tecnologias, bem como a recente medida de aumentar a escolaridade obrigatória para os 12 anos, e observar as suas implicações nas próximas avaliações do PISA.

Aspectos internacionais

O português é uma língua global com cerca de 220 milhões de falantes. Em Portugal existem cerca de 10 milhões de falantes [x] e em muitos países africanos o português é a língua oficial, mas coexistindo com muitas outras línguas nacionais (na sua maioria bantu, em Angola e Moçambique, Tétum, em Timor-Leste, e os Crioulos de Cabo Verde, Guiné-Bissau e São Tomé e Príncipe). Contudo, é de facto o Brasil que acolhe a maior parte da comunidade global de falantes do português, com 190 milhões de falantes nativos. Paralelamente ao tamanho da sua população, o Brasil está a contribuir para a projecção internacional cada vez maior da língua portuguesa, como resultado do seu desenvolvimento económico e da sua posição na cena internacional como uma das potências emergentes do século XXI.

Observa-se, portanto, um recente aumento do interesse pela língua portuguesa, especialmente nos países da América Latina, mas também em Macau e Espanha, por exemplo, sendo que a língua portuguesa é ensinada em muitos países do mundo[xi]. Diversas Câmaras de Comércio têm demonstrado interesse em oferecer aulas de português para potenciais investidores em Portugal, como foi o caso recente da Câmara Italiana, só para citar um exemplo entre muitos outros. As comunidades de emigrantes portugueses na Europa também promovem o ensino do português em vários países europeus.

Como consequência das explorações marítimas portuguesas, das descobertas geográficas e da implementação de novas rotas no comércio mundial, ainda no século XII, a língua portuguesa foi projectada durante séculos em todo o mundo como uma das línguas mais importantes para o comércio e para os negócios. É actualmente uma das 23 línguas oficiais da União Europeia e foi incluída em muitos projectos de investigação financiados pela Comissão Europeia com o objectivo de desenvolver recursos e tecnologias da linguagem. A língua portuguesa é também língua administrativa e de trabalho de 27 organizações internacionais, incluindo, por exemplo, a CPLP (Comunidade dos Países de Língua Oficial Portuguesa), a União Latina, o Mercosul e a FIFA (Federação Internacional de Futebol). Para além disso, nos últimos anos têm sido feitos esforços para incluir o português como uma das línguas oficiais das Nações Unidas.

Em conjunto com a sua progressiva projecção, a língua portuguesa enfrenta desafios nalguns contextos, quando se trata da sua posição como língua de comunicação internacional. Na América do Sul, com cerca de 190 milhões de falantes, o português co-existe com grandes comunidades de falantes de espanhol. Na Europa, o português tem pouco mais de 10 milhões de falantes. Na Ásia, o português é língua oficial somente em Timor-Leste e Macau. E em África, além do facto de muitas línguas nativas co-existirem com o português, o inglês e o francês são línguas com uma forte projecção nesse continente.

O português na internet

Uma visão geral dos dados estatísticos sobre a língua portuguesa mostra que é uma das línguas mais utilizadas na internet. De acordo com as últimas estimativas, o português é a quinta língua mais usada na internet, sendo ultraprassado apenas pelo inglês, chinês, espanhol e japonês[xii]. 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 em 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.[xiii]

Isto está de acordo com o boom de acesso à internet no Brasil, particularmente entre os jovens. Este país tem um dos maiores números de utilizadores de internet em todo o mundo (72 milhões)[xiv], e um questionário dos censos revelou que o número de utilizadores da internet com 10 anos ou mais deu um salto de 12 milhões desde 2008[xv]. Por sua vez Portugal tem cerca de 5 milhões de utilizadores de internet [xvi] e as estatísticas revelam que o número de subscritores de internet tem registado um notório aumento: em 2001 havia 466.813 assinantes, e as últimas estimativas indicam 1.898.026 assinantes. Estas revelam também que, em 2010, 54% dos lares portugueses 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 internet (92,7%), e que, em 2006, 95% das empresas com dez ou mais funcionários usavam computador, enquanto 84% utilizavam o email e 83% tinham acesso à internet[xvii].
Paralelamente ao esforço de assegurar a presença de institutos públicos, agências e serviços na internet, foi implementado em Portugal, em 2007, o Plano Nacional para a Promoção da Acessibilidade, em conjunto com legislação específica[xviii] orientada para a inclusão social através da Sociedade de Informação e para permitir o acesso a conteúdos na rede por parte de cidadãos com deficiência.

É assim claro o uso crescente do português 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, como nos *sites* da União Europeia ou do Mercosul, só para citar dois exemplos, embora devam continuar os esforços para que esta língua esteja presente noutras instituições onde ainda não é opção.

Leituras seleccionadas

Cardeira, Esperança, 2006, *O Essencial sobre a História do Português* Lisboa: Editorial Caminho.

Lewis, M. Paul (ed.), 2009. *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com.

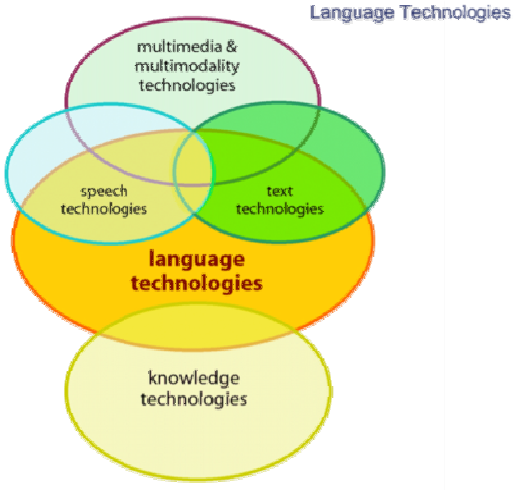
Pinto, Paulo Feytor, 2010, *Novo Acordo Ortográfico da Língua Portuguesa*. Lisboa: INCM.

Centro Virtual Camões(http://cvc.instituto-camoes.pt/conhecer/bases-tematicas.html)

Tecnologias da Linguagem para o português

Tecnologias da Linguagem

As Tecnologias da Linguagem correspondem a tecnologias da informação especializadas em lidar com a linguagem humana. Por esse motivo, são muitas vezes incluídas no grupo designado por Tecnologias da Linguagem Humana. A linguagem humana surge na forma falada e escrita. Enquanto a fala representa o meio mais antigo e natural de comunicação, os textos escritos, por seu lado, conseguem transmitir informação mais complexa e estruturar melhor o conhecimento humano. Hoje em dia, estes dois tipos de discurso podem ser reproduzidos por tecnologias de processamento e reconhecimento de fala e texto. Mas a linguagem também apresenta certos aspectos que são partilhados pela fala e pelo texto escrito, como os dicionários, a maior parte da gramática e o significado das frases. Assim, grande parte das tecnologias linguísticas não pode ser atribuída a uma categoria específica de fala ou texto. Entre estas tecnologias, encontram-se as que estabelecem ligações entre linguagem e conhecimento. A figura à direita ilustra o cenário das Tecnologias da Linguagem. No acto de comunicação, os falantes conciliam a linguagem com outros modos de comunicação e outros meios de informação. Combinamos o discurso com gestos e expressões faciais. Os textos digitais são acompanhados por imagens e sons. Os filmes podem conter linguagem sob as formas oral e escrita. Todas estas características contribuem para que as tecnologias responsáveis pelo processamento e reconhecimento de voz e texto se sobreponham e interajam com muitas outras tecnologias de modo que o processamento da comunicação intermodal e de documentos multimédia se torne mais fácil e acessível.



Arquitecturas de aplicações na área das Tecnologias da Linguagem

As aplicações mais usuais para o processamento da linguagem são constituídas por vários componentes que reflectem não só as diferentes etapas necessárias ao processamento da linguagem como também os diferentes módulos que a ferramenta terá de implementar. A figura à direita mostra, de um modo bastante simplificado, a arquitectura que pode ser encontrada num sistema de processamento de texto. Os três primeiros módulos ocupam-se da estrutura e do significado do texto de entrada (*input*):

- Pré-processamento: limpeza dos dados; remoção da formatação; detecção do idioma, etc.; identificação da estrutura das frases.
- Análise gramatical: detecção do verbo e dos seus complementos, modificadores, etc.
- Análise semântica: desambiguação (qual dos significados de "maçã" é o certo em determinado contexto?); resolução de anáforas e expressões referenciais, como "ela", "o carro", etc.; representação do significado da frase num modelo interpretável pela máquina.

Posteriormente, alguns módulos específicos podem executar outro tipo de operações, como a sumarização automática de um texto de entrada ou pesquisas em bancos de dados, entre outras. Nas secções seguintes apresentar-se-ão algumas **áreas centrais de aplicações**, tendo-se destacado alguns dos módulos das diferentes arquitecturas em cada secção. Também neste caso as arquitecturas

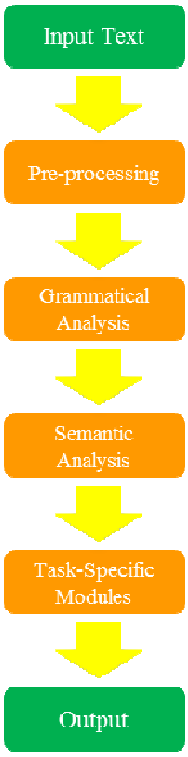


Figure 2: A Typical Text Processing Application Architecture

se encontram bastante simplificadas, uma vez que o principal objectivo consiste em ilustrar a complexidade das aplicações de Tecnologia da Linguagem de uma forma bastante geral e perceptível.

Após a discussão das áreas centrais de aplicações, descrever-se-á sumariamente o estado das Tecnologias da Linguagem na Educação e na Investigação, que será concluído com uma visão geral sobre antigos programas de financiamento. No final da secção, apresentar-se-á uma previsão feita por especialistas no que respeita à disponibilidade, maturidade e qualidade das ferramentas e dos recursos na área das Tecnologias da Linguagem, o que fornecerá uma boa visão sobre o estado da arte relativamente a este tipo de tecnologias em Portugal.

As ferramentas e os recursos mais importantes encontram-se sublinhados no texto, podendo também ser encontrados na tabela localizada no final do capítulo. As secções que retratam as áreas centrais das aplicações também apresentam uma visão geral das instituições no activo em Portugal e no Brasil, para as respectivas áreas.

Áreas centrais de aplicações

Correctores ortográficos e sintácticos

Qualquer pessoa que use uma ferramenta de processamento de texto, como o MS Word, depara-se com uma componente de verificação ortográfica (correctores) que indica os erros ortográficos e propõe correcções. Hoje em dia, quarenta anos após o primeiro programa de correcção ortográfica, concebido por Ralph Gorin, os correctores tornaram-se extremamente sofisticados, não se limitando a comparar determinadas palavras com aquelas que têm presentes nos dicionários. Além da existência de algoritmos dependentes da linguagem, que permitem o tratamento da morfologia (formação do plural, por exemplo), alguns correctores incluem também módulos que os tornam capazes de reconhecer erros relacionados com a sintaxe, tais como um verbo em falta ou um verbo que não concorda com o sujeito em pessoa e número, como em "Elas *escreve uma carta". Apesar disso, a maior parte dos correctores ortográficos (incluindo o MS Word) não encontrará quaisquer erros nos versos seguintes, pertencentes a um poema de Jerrold H. Zar (1992):

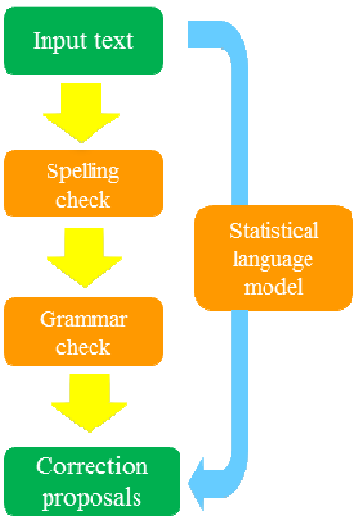


Figure 3: Language Checking (left: rule-based; right: statistical)

Eye have a spelling chequer,

It came with my Pea Sea.

It plane lee marks four my revue

Miss Steaks I can knot sea.

Para lidar com este tipo de erros, é necessária uma análise cuidada do contexto, como se pode demonstrar nos seguintes exemplos para o português:

Fizemos jogos tradicionais, incluindo o jogo do pião.

Fizemos jogos tradicionais, incluindo o jogo do peão.

Nestes casos, é exigida a formulação de regras gramaticais específicas da língua (que implica um elevado grau de especialização e trabalho manual) ou o uso do chamado modelo estatístico da língua. Estes modelos calculam a probabilidade de uma determinada palavra ocorrer num determinado contexto (isto é, tem-se em conta as palavras anteriores e seguintes). 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 da língua pode ser automaticamente obtido através do recurso a uma grande quantidade de dados correctos da língua (i.e., um *corpus*). Até agora, estas abordagens têm sido quase sempre desenvolvidas e avaliadas com dados para o inglês. No entanto, tais aplicações não podem ser directamente transferidas para o português, dada a sua riqueza a nível de flexão verbal, por exemplo.

O uso de correctores ortográficos não se limita às ferramentas de processamento de texto, sendo também aplicado em sistemas de apoio ao autor. Para acompanhar o aumento de produtos técnicos, os documentos de carácter mais terminológico também têm sofrido um grande acréscimo nas últimas décadas. Temendo as reclamações dos clientes devido à utilização errada dos produtos e os danos resultantes de uma possível má interpretação dos manuais de instrução, as empresas começaram a concentrar-se cada vez mais na qualidade técnica da documentação, visando, ao mesmo tempo, o mercado internacional. Os avanços na área do processamento da linguagem natural levaram ao desenvolvimento de aplicações de apoio ao autor, que auxiliam o redactor de documentação técnica no uso de vocabulário e de estruturas de frases, de acordo com certas regras e restrições de terminologia.

Além do MS Word, existem, para o português, outras ferramentas de correcção ortográfica. Em Portugal, a empresa Priberam criou o FLIP, um software que disponibiliza vários produtos na área da verificação ortográfica e sintáctica para o português (tanto português europeu como português do Brasil) e o espanhol. O CoGrOO, para o Open Office, é um corrector gramatical para o português do Brasil. Também para esta variante do português, o NILC (Núcleo Interinstitucional de Lingüística Computacional) desenvolveu o ReGra, que está disponível como parte integrante do MS Word e do processador de texto REDATOR.

Além dos correctores 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 correcção automática de pesquisas enviadas para motores de busca da internet, como é o caso das sugestões do Google "Você quis dizer ...".

Pesquisa na internet

A pesquisa na internet, em intranets ou em bibliotecas digitais é provavelmente a Tecnologia da Linguagem mais utilizada e também a menos desenvolvida nos dias de hoje.

O motor de pesquisa Google, que surgiu em 1998, recebe actualmente cerca de 80% das pesquisas que se fazem na internet em todo o mundo. O verbo *googlar* até tem uma entrada no dicionário *online* da Porto Editora. Nem a interface de pesquisa nem a apresentação dos resultados obtidos sofreram alterações significativas desde a primeira versão deste motor de pesquisa. Na actual versão, o Google oferece uma correcção ortográfica para as

palavras com erros ortográficos. A sua capacidade de pesquisa semântica, que, desde 2009, se encontra incorporada no seu algoritmo, permite-lhe melhorar a precisão da mesma, analisando o significado dos termos da consulta no seu contexto. A história de sucesso do Google mostra que, na posse de um conjunto de dados e de técnicas eficientes para a indexação desses dados, uma abordagem essencialmente baseada em estatística pode levar a resultados satisfatórios.

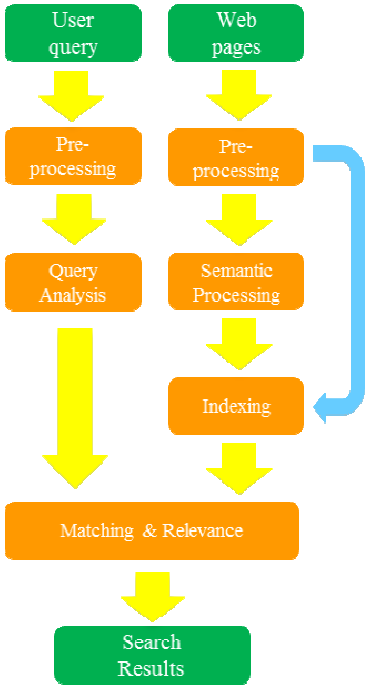


Figure 4: Web Search Architecture

No entanto, para um pedido de informação mais elaborado, é essencial integrar conhecimentos de linguística mais profundos. As experiências realizadas com recurso a *thesauri* e bases de dados ontológicas (como a WordNet) criados num formato interpretado pelo computador têm apresentado resultados satisfatórios, como a possibilidade de encontrar uma página a partir de sinónimos dos termos de pesquisa (por exemplo, "energia atómica", "energia nuclear" e "centrais nucleares") ou até a partir de termos mais vagamente relacionados. Nestes casos, e para o português, será crucial o uso de recursos como as WordNets MWN.PT e Word-Net.PT.

A próxima geração de motores de pesquisa terá de incluir Tecnologias da Linguagem muito mais sofisticadas. Se em vez de uma lista de palavras-chave a pesquisa 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 sintáctico e semântico, como também a disponibilização de um índice que permita uma recuperação rápida dos documentos pertinentes. Por exemplo, suponhamos que um utilizador faz a seguinte pesquisa: "Dá-me me uma lista de todas as empresas que foram compradas por outras empresas nos últimos cinco anos". Para uma resposta satisfatória, é necessário proceder-se a uma análise sintáctica da frase (*parser*) para observar a sua estrutura gramatical e determinar que o que o utilizador está à procura é de empresas que foram compradas e não de empresas que compraram outras. Além disso, é igualmente preciso processar a expressão "últimos cinco anos" para descobrir quais os anos a que se refere exactamente.

Finalmente, é necessário que a pesquisa processada seja comparada com uma grande quantidade de dados não estruturados com o objectivo de encontrar parte ou partes da informação que o utilizador está a procurar. Este processo é normalmente referido como recuperação de informação (*information retrieval*) e envolve tarefas de pesquisa em documentos considerados relevantes. No caso da pesquisa acima referida, para obter 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 uma ferramenta de reconhecimento de entidades nomeadas.

Ainda mais exigente é a tentativa de fazer corresponder uma pesquisa a documentos escritos em línguas diferentes. Para a recuperação de informação interlínguas, temos de traduzir automaticamente a pesquisa para todas as línguas de origem possíveis e transferir a informação recolhida de volta para a língua-alvo. A crescente percentagem de dados disponíveis em formatos não-textuais leva à procura de serviços que permitam a recuperação de informação multimédia, isto é, a pesquisa de informação em imagens, áudio e vídeo. Para ficheiros de áudio e vídeo, esta tarefa envolve um módulo de reconhecimento de voz, que tem como função converter o conteúdo da fala num formato de

texto ou numa representação fonética para os quais se possa fazer uma correspondência com as pesquisas feitas pelo utilizador.

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

A situação no Brasil é um pouco diferente. Há exemplos de motores de pesquisa direccionados apenas para *sites* brasileiros (como o Achei ou o Giga Busca), mas são em menor número do que em Portugal, e a sua cobertura e o seu alcance são bastante limitados. Por este motivo, o Google é também o motor de pesquisa dominante no Brasil.

Interacção de fala

A tecnologia de interacção de fala é a base para a criação de interfaces que permitem ao utilizador interagir com máquinas que utilizam a linguagem falada em vez de, por exemplo, um monitor, um teclado e um rato. Actualmente, estas interfaces de utilização de voz podem ser parcial ou totalmente automatizadas e são geralmente utilizadas por empresas para oferta de serviços aos seus clientes, empregados ou associados, via telefone. Os negócios na área da banca, logística, transportes públicos e telecomunicações são dos que mais fortemente apostam neste tipo de aplicações. A tecnologia de interacção de fala apresenta ainda outros tipos de utilizações, tais como interfaces para determinados 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 é o caso dos *smartphones*.

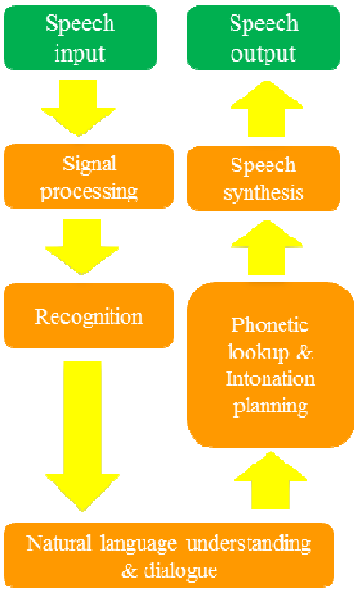


Figure 5: Simple Speech-based Dialogue Architecture

Na sua essência, a interacção de fala compreende quatro diferentes tecnologias:

- O reconhecimento automático de fala é responsável por determinar que palavras foram efectivamente pronunciadas numa determinada sequência de sons produzidos por um utilizador.
- A análise sintáctica e a interpretação semântica analisam a estrutura sintáctica do enunciado produzido pelo utilizador e interpretam-no de acordo com o objectivo do respectivo sistema.
- A gestão do diálogo é necessária para determinar, por parte do sistema com o qual o utilizador interage, que acção deve ser tomada tendo em conta o *input* do utilizador e a funcionalidade do próprio sistema.
- A tecnologia de síntese de voz (Texto-Som) é utilizada para transformar as palavras que constituem o texto de um enunciado em sons, que serão fornecidos ao utilizador.

Um dos grandes desafios consiste em ter um sistema de reconhecimento automático de fala que reconheça as palavras pronunciadas por um utilizador com a maior exactidão possível. Para tal, são necessários alguns requisitos, tais como a restrição do conjunto de enunciados possíveis a um conjunto limitado de palavras-chave, ou a criação manual de modelos de linguagem que cubram uma grande variedade de expressões de língua natural. Enquanto a primeira opção resulta em utilizações bastante rígidas e inflexíveis das interfaces de utilização de voz, tendo como consequência uma fraca aceitação por parte dos utilizadores, a criação, afinação e manutenção de modelos de linguagem podem elevar significativamente os custos. No entanto, as interfaces que empregam modelos de linguagem e que permitem inicialmente a um utilizador expressar a sua intenção de forma flexível – evocada, por exemplo, pela pergunta *"Como posso ajudá-lo"* – mostram quer uma taxa mais elevada de automação quer uma maior aceitação por parte do utilizador, podendo, assim, ser consideradas mais vantajosas do que as abordagens dirigidas para o diálogo directo, que são menos flexíveis.

Quando se trata de produzir o *output* por parte de uma interface de utilização de voz, as empresas tendem a fazer amplo uso de enunciados pré-gravados por locutores profissionais. Para enunciados estáticos, em que o texto não depende de contextos particulares nem de dados pessoais do utilizador, tal aplicação resultará numa experiência enriquecedora. No entanto, quanto mais dinâmico for o conteúdo de um enunciado que o sintetizador tem de considerar, mais hipóteses há de que os resultados de *output* apresentem uma prosódia pobre, resultante da concatenação de arquivos de áudio individuais. Em contrapartida, e tendo em conta que podem ser optimizados, os actuais sistemas de sintetizadores de fala mostram uma considerável superioridade em relação à naturalidade prosódica de enunciados dinâmicos.

Relativamente ao mercado das tecnologias de interacção de fala, a última década tem sido caracterizada quer por uma forte padronização das interfaces para os diferentes componentes tecnológicos, quer pelo estabelecimento de padrões para a criação de artefactos específicos de *software* para uma determinada aplicação. Houve também uma forte consolidação do mercado nos últimos 10 anos, em particular nas áreas de reconhecimento e síntese de fala.
Os mercados nacionais dos países do G20 – ou seja, os países economicamente mais fortes e com uma população considerável – são dominados por menos de cinco aplicações em todo o mundo, com destaque para a proeminência, na Europa, do *Nuance* e do *Loquendo*.

No mercado português, existem algumas pequenas empresas, como a SVOX e a Voice Interaction, tendo esta última a particularidade de fazer reconhecimento e síntese de fala não apenas para o português europeu e do Brasil, mas também para as variedades africanas do português.

No que respeita à tecnologia de gestão de diálogo e de *know-how*, DigA é a única aplicação completa, construída especificamente para o português europeu: é de domínio público, mas não está disponível com 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 oral, o gestor de diálogo é o único módulo que pode ser usado para qualquer língua. Os outros módulos existem – embora geralmente não sejam de livre acesso nem estejam disponíveis com

código aberto –, mas a tarefa de adaptação para uma determinada língua exige muito tempo e esforço humano.

Por último, no domínio da interacção de voz, ainda não existe um verdadeiro mercado para as principais tecnologias linguísticas para a análise sintáctica e semântica dos enunciados.

Olhando para lá do estado actual da tecnologia, podem adivinhar-se mudanças significativas devido à propagação de *smartphones* como uma nova plataforma para o gerenciamento de relações com clientes – além dos canais de internet, telefone e e-mail. Esta tendência também afectará o emprego da tecnologia nas aplicações de interacção de voz. Por um lado, a procura de aplicações de utilização de voz para chamadas telefónicas irá diminuir a longo prazo. Por outro lado, o uso da língua falada como uma modalidade de comunicação amigável para *smartphones* irá ganhar uma importância significativa. Esta tendência é reforçada pela observável melhoria da precisão de reconhecimento do discurso independente do falante para serviços de ditado de voz, que são já oferecidos como serviços centralizados para utilizadores de *smartphones*. Dado este *outsource* da tarefa de reconhecimento relativamente à infra-estrutura das aplicações, o emprego de aplicações específicas de tecnologias de base linguística irá, supostamente, ganhar importância em comparação com a situação de hoje.

Quanto à tecnologia de gestão de diálogo, será fundamental que esta aumente o seu raio de acção e apoie cenários de interacção multimodal (como aqueles criados pelo uso de *smartphones*), bem como canais de interfaces de múltiplos utilizadores, a partir de um modelo comum de domínio de comportamento específico de interacção. Além da investigação em curso sobre a optimização do reconhecimento e da síntese de fala, as áreas de personalização de domínios específicos de tecnologias de base linguística e de gestão de diálogo parecem ser as mais relevantes no que respeita a áreas de transformação entre a investigação aplicada e produção industrial.

Tradução automática

A ideia de usar computadores para a tradução das línguas naturais surgiu em 1946 pela mão de A. D. Booth. Seguidamente, nos anos 50 e, novamente, no início dos anos 80, procedeu-se a financiamentos substanciais nesta área de investigação. Contudo, a tradução automática ainda falha em cumprir as altas expectativas que gerou nos primeiros anos de investigação.

No seu nível mais básico, a tradução automática apenas substitui as palavras numa língua natural por outras palavras noutra língua natural. Isto poderá ser útil em domínios muito restritos e linguagens terminológicas, 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 da língua-alvo. Neste caso, a maior dificuldade reside no facto de a linguagem humana ser ambígua. A desambiguação de palavras revela-se, assim, um grande desafio a vários níveis. Por exemplo, a nível lexical, "banco" apresenta, pelo menos, dois significados: "peça de mobiliário para as pessoas se sentarem" e "edifício onde se realizam operações financeiras".

O rapaz viu a rapariga no banco.

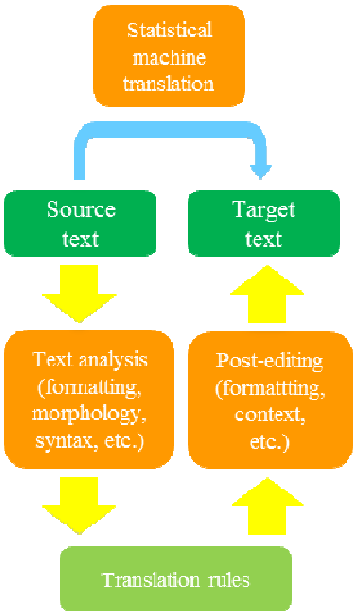


Figure 6: Machine translation (top: statistical; bottom: rule-based)

A ambiguidade sintáctica também apresenta grandes desafios, como se pode observar pelos exemplos seguintes, em que as frases são estruturalmente idênticas, mas uma apresenta ambiguidade e a outra não.

O polícia viu o homem com o telescópio.

(O polícia viu o homem através do telescópio / O polícia viu o homem que levava o telescópio.)

O polícia viu o homem com o revólver.

(O polícia viu o homem que levava o revólver)

Uma das formas de abordar esta tarefa de desambiguação consiste na utilização de ferramentas baseadas em regras linguísticas. Para traduções entre línguas aproximadas, a tradução directa em casos como os exemplos acima é exequível. Mas, muitas vezes, os sistemas baseados em regras (ou conduzidos para o conhecimento) analisam o texto de entrada e criam um texto intermediário (representação simbólica) a partir do qual o texto na língua-alvo é gerado. O sucesso destes métodos está fortemente dependente da disponibilidade não só de léxicos extensivos com informação morfológica, sintáctica e semântica, como também de grandes conjuntos de regras gramaticais concebidas cuidadosamente por um linguista especializado.

A partir dos finais dos anos 80, altura em que os recursos computacionais aumentaram e se tornaram menos dispendiosos, começou a surgir um maior interesse na criação de modelos estatísticos para a TA. Os parâmetros destes modelos derivam da análise de *corpora* de textos bilingues, como o *corpus* paralelo Europarl, que contém as actas do Parlamento Europeu em onze línguas europeias. Com dados suficientes, a TA baseada em dados estatísticos funciona bem o suficiente para produzir um significado aproximado para um texto numa língua estrangeira. Contudo, ao contrário dos sistemas conduzidos para o conhecimento, a TA baseada em estatística pode gerar muitas vezes resultados com erros gramaticais. Por outro lado, além da vantagem de ser necessário um menor esforço humano para escrever a gramática, a TA baseada em estatística pode também cobrir particularidades da língua que vão falhar em sistemas baseados no conhecimento, como, por exemplo, as expressões idiomáticas.

Devido ao facto de os pontos fortes e os pontos fracos destes dois tipos de abordagem da TA serem complementares, actualmente, os investigadores são unânimes em desenvolver abordagens híbridas, combinando metodologias de ambas. Tal procedimento pode ser feito de várias maneiras. Uma consiste em utilizar tanto o modelo do conhecimento como o dos dados e ter um módulo de selecção que decida o melhor *output* para cada frase. No entanto, para frases mais longas, nenhum resultado será perfeito. A melhor solução será a de combinar as melhores partes de cada frase resultantes de múltiplos *outputs*, o que poderá ser bastante complexo, uma vez que as partes correspondentes a múltiplas alternativas nem sempre são claras e precisam de ser alinhadas.

No caso do português, a falta de um mecanismo eficaz de desambiguação de palavras é uma das razões principais para que os resultados dos sistemas de tradução automática existentes sejam muitas vezes insatisfatórios.

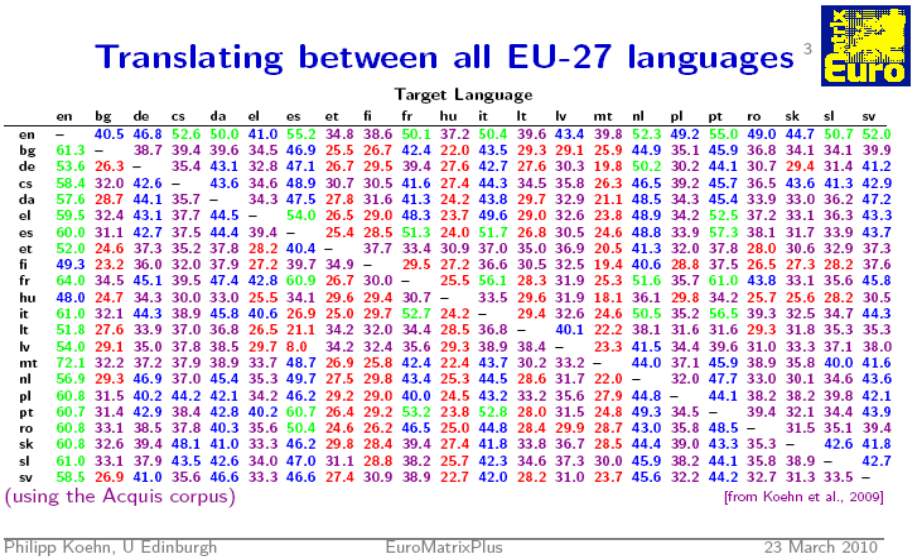
Além disso, enquanto línguas como o alemão, por exemplo, formam compostos constituídos por uma única palavra, a tendência no português é para escrever os compostos como sintagmas, ou seja, separar as palavras que formam uma unidade lexical. No que ao português diz respeito, este poderá ser um desafio específico para os tradutores automáticos.

Alguns dos mais importantes sistemas de tradução automática baseados em regras, como o LOGOS, o Apertium e o SYSTRAN, estão disponíveis para o português. Apesar de haver uma investigação significativa nesta área da tecnologia, tanto no contexto nacional como internacional, os sistemas híbridos e baseados em regras têm sido, até agora, menos bem-sucedidos no ramo dos negócios do que no da investigação.

Ao fornecer uma boa adaptação relativamente a terminologias específicas e integração de *workflows*, o uso de tradutores automáticos pode aumentar significativamente a produtividade. Alguns sistemas especiais para apoio à tradução interactiva foram desenvolvidos, por exemplo, pela Siemens. Alguns portais da língua, como o da Volkswagen, proporcionam o acesso a dicionários, terminologias específicas da empresa, memória de tradução e apoio para tradução automática.

A qualidade dos sistemas de tradução automática ainda apresenta um grande potencial de optimização. De entre os desafios existentes, destacam-se a adaptabilidade dos recursos da língua a um determinado domínio e a sua integração em *workflows* com bases de dados terminológicas e memórias para tradução. Além disso, a maioria dos actuais sistemas é direccionada para o inglês e suporta poucas línguas, sendo necessária a tradução do português e para o português, o que não só leva a deficiências no *workflow* total de tradução, como força os utilizadores de tradutores automáticos a aprender diferentes ferramentas de codificação lexical para diferentes sistemas.

As campanhas de avaliação permitem a comparação da qualidade dos sistemas de tradução automática, das diferentes abordagens e do estatuto dos sistemas de tradução para as diferentes línguas. O quadro em baixo, apresentado no âmbito do projecto CE Euromatrix+, mostra a comparação dos resultados obtidos para as 22 línguas oficiais da União Europeia (com excepção do gaélico irlandês) em termos de classificação BLEU[xix]..



	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	–

Os melhores resultados (a azul e a verde) foram conseguidos tanto por línguas que beneficiam de consideráveis esforços de investigação no âmbito de programas coordenados (a partir da existência de *corpora* paralelos), como por línguas que apresentam alguma semelhança com outras línguas (como o inglês, o francês, o holandês, o espanhol, o português ou o alemão). Os piores resultados (a vermelho) dizem respeito a línguas que não beneficiaram de esforços semelhantes ou que são muito diferentes de outras línguas (como o húngaro, o maltês ou o finlandês).

"Os bastidores" das Tecnologias da Linguagem

A construção de aplicações na área das Tecnologias da Linguagem envolve uma série de tarefas que nem sempre são visíveis no nível de interacção com o utilizador, mas que fornecem funcionalidades significativas "escondidas" pelo sistema. Por conseguinte, tais tarefas constituem pontos cruciais de investigação, tendo-se inclusivamente tornado em subdisciplinas da Linguística Computacional no meio académico.

A área da criação de sistemas de resposta-a-perguntas, por exemplo, transformou-se numa das mais activas na investigação, tendo-se procedido à construção de *corpora* anotados. A ideia é passar de uma pesquisa baseada em palavras-chave (à qual o motor de pesquisa responde 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 (por exemplo, pergunta: "Com que idade Neil Armstrong pisou a Lua?"; resposta: "38 anos"). Enquanto isto está evidentemente relacionado com o que foi acima referido sobre as pesquisas na internet, o termo resposta-a-perguntas tem funcionado, hoje em dia, como um termo geral para questões de pesquisa, como, por exemplo, que tipos de perguntas devem ser considerados e como é que devem ser tratados, como é que um conjunto de documentos potencialmente detentores da resposta devem ser analisados e comparados (será que dão respostas antagónicas?) e como é que uma informação específica – a resposta

– pode ser fielmente extraída de um documento sem ignorar inadvertidamente o contexto.

Esta questão está, por sua vez, relacionada com a tarefa de extracção de informação, uma área que foi extremamente popular e influente na "era da estatística", em Linguística Computacional, no início da década de 1990. A extracção de informação tem como objectivo identificar conteúdos específicos de informação em determinados tipos de documentos (por exemplo, a detecção de figuras-chave na aquisição de empresas tal qual foi relatado nos jornais). Um outro cenário que tem sido trabalhado diz respeito a relatórios sobre incidentes terroristas, em que o problema reside no mapeamento de um texto a um modelo que indique o agressor, o alvo, a hora, o local e os resultados do incidente. A principal característica da extracção de informação consiste, assim, no preenchimento de modelos de domínios específicos. Esta tarefa representa, deste modo, um bom exemplo do que se passa nos "bastidores" das tecnologias da linguagem e, por constituir uma área de investigação bem delimitada e pelos fins práticos que possui, necessita de ser enquadrada numa aplicação adequada.

A sumarização e a geração automática de textos constituem duas áreas de contacto, mas que, muitas vezes, funcionam como aplicações individuais ou de suporte. A sumarização refere-se, naturalmente, à tarefa de encurtar um texto longo e vem, por exemplo, como uma funcionalidade do MS Word. Esta aplicação funciona, em grande parte, com base em métodos estatísticos: identifica primeiramente palavras "importantes" num texto (que são, 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, selecciona 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 utilizado, a sumarização corresponde ao processo de extracção 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, a que alguns investigadores dedicam particular atenção, consiste em sintetizar efectivamente frases novas, ou seja, construir um resumo com construções sintácticas que não aparecem no texto de origem. Esta tarefa exige uma compreensão muito mais aprofundada do texto e, por esse motivo, é muito menos sólida. Convém realçar que um gerador automático de texto não representa, na maior parte dos casos, uma aplicação individual, encontrando-se incorporado numa aplicação mais ampla (como é o caso dos sistemas de informação de clínicas médicas, nos quais os dados dos doentes são recolhidos, armazenados e processados e em que a geração automática de relatórios é apenas uma das suas muitas funcionalidades).

Nestas áreas, a investigação tem recaído mais sobre a língua inglesa, pelo que o desenvolvimento de sistemas de resposta-a-perguntas, extracção de informação e sumarização automática têm sido objecto, desde a década de 90, de numerosos concursos para atribuição de financiamento, como os organizados pela DARPA/NIST, nos Estados Unidos. Esta investigação tem contribuído significativamente para o avanço do estado da arte. No entanto, a língua-alvo tem sido sempre o inglês. Alguns concursos têm acrescentado opções multilingues, mas o português, como muitas outras línguas, não tem recebido apoio suficiente. Por conseguinte, não existe praticamente *corpora* anotados ou outros recursos necessários para o desenvolvimento destas aplicações. Os

sistemas de sumarização automática que utilizem simplesmente métodos estatísticos são, em grande parte, independentes da língua, e, nestes casos, encontram-se disponíveis alguns modelos protótipos. No entanto, já existe especificamente para o português uma ferramenta de sumarização que utiliza métodos estatísticos mas com base na ideia principal do texto. 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 (como as "gramáticas generativas"). Mas, mais uma vez, as aplicações disponíveis estão direccionadas para o inglês, não havendo, nesta área, ferramentas disponíveis para o português. De igual modo, podemos encontrar apenas um número muito limitado de sistemas de resposta-a-perguntas para o português.

As Tecnologias da Linguagem na educação

A área das Tecnologias da Linguagem destaca-se pela sua interdisciplinaridade, envolvendo uma grande variedade de domínios científicos, como a Linguística e as Ciências Computacionais, a Estatística, a Engenharia e a Psicolinguística, entre muitos outros. No que diz respeito ao Ensino Superior, Portugal apresenta uma oferta aceitável nesta área, com cursos relevantes, como Tradução, Ciências da Linguagem ou Ciências Computacionais.

A área das Tecnologias da Linguagem foi desenvolvida em muitas universidades, quer na educação (licenciaturas, mestrados e doutoramentos) quer em centros de investigação. Na Universidade de Lisboa, a par de diversos cursos com diferentes níveis de ensino (incluindo um *minor* em Processamento de Linguagem Natural e um programa de mestrado e de doutoramento em Ciências Cognitivas), existem importantes centros de investigação dedicados às Tecnologias da Linguagem. O Grupo de Fala e Linguagem Natural (NLX), do Departamento de Informática da Faculdade de Ciências, é, a nível nacional, a equipa líder no processamento computacional do português, disponibilizando *online* um conjunto abrangente de serviços de processamento linguístico (LX-Center). 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 dialectal ou mesmo da história da língua), tendo construído um *corpus* de grande escala, de que resultou a criação de recursos mais pequenos e específicos, disponíveis *online*.

O Instituto Superior Técnico (IST), em Lisboa, além de oferecer cursos em Tecnologias da Linguagem, também apresenta 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 do IST e o seu Laboratório de Sistemas de Língua Falada (L2f) é o líder nacional na produção de sistemas de reconhecimento e síntese de fala.

A Universidade Nova de Lisboa também tem cursos no campo das Tecnologias da Linguagem, bem como unidades de investigação, nomeadamente o Centro de Investigação em Tecnologias de Informação (CITI) e o Centro de Linguística (CLUNL).

Ainda em Lisboa, existe o ILTEC, um instituto dedicado à linguística teórica e computacional. No resto do país, também existem universidades que oferecem cursos na área das

Tecnologias da Linguagem e que também acolhem centros de investigação, como o Centro de Investigação em Tecnologias de Informação, na Universidade de Évora; o Centro de Estudos de Linguística Geral e Aplicada (CELGA), na Universidade de Coimbra; o Centro de Tecnologia da Linguagem Humana e Bioinformática (HULTIG), na Universidade da Beira Interior; o Centro de Linguística (CLUP) e o Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), na Universidade do Porto ou o Centro de Estudos Humanísticos (CEHUM), na Universidade do Minho. A Universidade do Algarve tem cooperado com o programa europeu Erasmus para a realização de mestrados na área do Processamento das Línguas Naturais e Tecnologia da Linguagem Humana.

No Brasil, também se tem assistido a uma actividade considerável na área das Tecnologias da Linguagem (tanto no ensino como na investigação), que se concentra sobretudo nas áreas urbanas de São Paulo, Rio de Janeiro e Porto Alegre. No entanto, nesta área, os cursos têm sido ministrados mais a nível de pós-graduações (mestrados e doutoramentos) do que de licenciatura.

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

Programas no âmbito das Tecnologias da Linguagem

A actividade na área das Tecnologias da Linguagem em Portugal pode ser avaliada através dos projectos, programas ou iniciativas levados a cabo nas últimas décadas. Um dos primeiros importantes programas nesta área foi o EUROTRA, um ambicioso projecto de tradução automática criado e financiado pela Comissão Europeia desde o final dos anos 70 e que durou até 1994. Portugal entrou neste projecto em 1986 através do Instituto de Linguística Teórica e Computacional (ILTEC), criado especificamente para este propósito. Este projecto teve um impacto duradouro na indústria das línguas a nível europeu. O EUROTRA foi um ponto de partida importante para o desenvolvimento contínuo de actividades no âmbito das Tecnologias da Linguagem em Portugal e para a criação de uma comunidade portuguesa de investigadores nesta área.

O projecto LE-PAROLE, desenvolvido no final dos anos 90, com a participação do CLUL e do INESC, foi outro projecto-chave europeu na área das Tecnologias da Linguagem que envolveu a língua portuguesa. Dos seus resultados finais, destaca-se a construção de *corpora* e léxicos de acordo com modelos integrados de constituição e descrição de materiais, em que se usam ferramentas comuns, o que permite facilitar as ligações multilingues e dar resposta 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 inclui um *subcorpus* anotado de 250 000 palavras. O léxico de cada língua é composto por 20 000 entradas, com informação sintáctica e morfossintáctica.
Em Portugal, este *corpus* foi alargado e enriquecido com o projecto TAGSHARE, realizado na Universidade de Lisboa pelo Departamento de Informática (NLX) e pelo Centro de Linguística

(CLUL), em 2005. Este projecto permitiu o desenvolvimento de um conjunto de recursos linguísticos e de ferramentas que permitem melhorar o processamento computacional do português. Como resultado final, obteve-se um *corpus* de 1 milhão de palavras linguisticamente anotadas e manualmente revistas por especialistas – o *corpus* CINTIL –, bem como todo um conjunto de ferramentas para *tokenização*, anotação de categoria morfossintáctica (POS), flexão, lematização, reconhecimento de multipalavras, reconhecimento de entidades nomeadas, etc. Os sistemas de anotação desenvolvidos no âmbito deste projecto tornaram-se, inclusivamente, em modelos para o português no campo das Tecnologias da Linguagem, sendo utilizados no Corpus de Referência do Português Contemporâneo (CRPC).

O Corpus de Extractos de Textos Electrónicos MCT/Público (CETEMPúblico) é um *corpus* com cerca de 180 milhões de palavras provenientes de textos de um jornal diário português lançado em 2000. Pretende, sobretudo, dar apoio ao desenvolvimento de ferramentas de processamento para o português, as quais, para a sua construção e avaliação, necessitam de textos em bruto. Este *corpus* foi criado no âmbito do projecto Processamento Computacional do Português, ao abrigo de um protocolo entre o Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) e o jornal *Público*. Posteriormente, este projecto evoluiu para a Linguateca, um projecto a longo prazo virado para as Tecnologias da Linguagem do português.

Do ponto de vista comercial, vale a pena destacar a presença em Portugal, desde 2005, do International Microsoft Language Development Center, que tem contribuído fortemente para o desenvolvimento da indústria das Tecnologias da Linguagem no nosso país.
Mais recentemente, algumas instituições portuguesas e brasileiras têm participado no projecto CLARIN (em curso), que tem como objectivo a criação de uma infra-estrutura europeia de investigação de recursos da língua e de tecnologia que seja integrada e interoperável.

No Brasil, também têm sido realizados esforços na área das Tecnologias da Linguagem para o português. Como exemplos, pode referir-se a criação do Banco de Português, no início dos anos 90, pela Pontifícia Universidade Católica de São Paulo, no âmbito do projecto DIRECT. Desde a sua criação, o Banco de Português tem sido uma importante fonte de dados para diversos estudos baseados em *corpus*. Vale também a pena referir o *corpus* Summit, construído para dar apoio a estudos de sumarização automática e de fenómenos anafóricos e de relações retóricas para o português. Este recurso foi desenvolvido no âmbito do projecto PLN-BR, pelo Núcleo Interinstitucional de Lingüística Computacional (NILC), levado a cabo pela Universidade de São Paulo e por um conjunto de investigadores de outras instituições.

Estes são apenas alguns exemplos de projectos, programas e iniciativas na área das Tecnologias da Linguagem para a língua portuguesa. Apesar de constituírem uma evolução positiva para a o português nos últimos anos, o facto é que, mesmo para línguas mais estudadas e para as quais o desenvolvimento de recursos linguísticos e tecnológicos se encontra numa fase muito mais avançada, existe ainda uma grande lacuna no que respeita à actividade das Tecnologias da Linguagem.

Comparado com o nível de financiamento para a área das Tecnologias da Linguagem nos Estados Unidos, o apoio para esta

área em Portugal e noutros países europeus 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 projectos em Tecnologias da Linguagem é particularmente difícil, uma vez que as propostas nesta área são submetidas e avaliadas em programas de Engenharia Electrotécnica, nos quais têm de competir com centenas de propostas de projectos sobre assuntos completamente diferentes. Além da FCT, a Fundação Calouste Gulbenkian também financia, ocasionalmente, projectos na área das Tecnologias da Linguagem.

No Brasil, o financiamento para a investigação em geral, e para as actividades em Tecnologias da Linguagem em particular, vem sobretudo de agências governamentais. O Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), o São Paulo Research Foundation (FAPESP), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a Pesquisa e Financiamento e Projetos (FINEP) são as quatro principais instituições de financiamento neste país. Algumas participaram inclusivamente em programas de financiamento conjunto com algumas universidades. Por exemplo, a FAPESP e o Microsoft Research Center formaram recentemente uma parceria para o financiamento de projectos socialmente relevantes no Estado de São Paulo, que incluiu, entre outros, o PorSimples, um projecto na área das Tecnologias da Linguagem que tem como objectivo a simplificação de textos de português para auxiliar leitores pouco alfabetizados a processar textos da internet.

Disponibilidade de ferramentas e recursos para o português

A tabela seguinte fornece uma visão genérica relativamente à situação actual das Tecnologias da Linguagem para o português. A avaliação das ferramentas e dos recursos existentes baseia-se numa análise informada de vários especialistas na área, considerando os seguintes critérios (cada um variando de 0 a 6).

1. **Quantidade:** A ferramenta/o recurso existem para a língua em questão? Quanto maior o número de ferramentas/recursos, melhor será a classificação.
   - 0: não há qualquer ferramenta/recurso
   - 6: muitas ferramentas/recursos; grande variedade
2. **Disponibilidade:** As ferramentas/os recursos estão acessíveis, ou seja, existem em código aberto, com utilização livre para qualquer plataforma, ou estão apenas disponíveis por um preço elevado ou em condições muito restritas?
   - 0: praticamente todos os recursos/ferramentas estão apenas disponíveis por um preço elevado
   - 6: uma grande parte de ferramentas/recursos é livre, estando disponível sob licenças *Open Source* ou *Creative Commons*, que permitem reutilização e adaptação para propósitos diferentes
3. **Qualidade** - Como se comportam os critérios de desempenho das ferramentas e os indicadores de qualidade dos recursos relativamente às melhores ferramentas, às aplicações ou aos recursos que se encontram disponíveis? As

ferramentas/os recursos encontram-se actualizados e são activamente mantidos?

- 0: recurso/ferramenta experimental
- 6: ferramenta de grande qualidade, recurso com anotação manual

4 **Cobertura:** Até que ponto as melhores ferramentas incluem critérios de cobertura (estilos, géneros, tipos de texto, fenómenos linguísticos, tipos de *input*/*output*, número de línguas admitidas por um sistema de tradução automática, etc.)? Até que ponto os recursos são representativos da língua-alvo ou de sub-línguas?

- 0: recurso ou ferramenta com fins específicos, casos específicos, muito pouca cobertura, de uso direccionado apenas para situações muito específicas, casos de uso particulares
- 6: recurso com muito boa cobertura, ferramenta muito robusta, aplicável a uma larga variedade de casos, admite muitas línguas

5 **Maturidade:** O recurso/a ferramenta podem ser considerados maduros, estáveis e prontos para o mercado? Os melhores recursos/ferramentas disponíveis estão prontos a ser imediatamente usados ou têm de ser adaptados? O desempenho desta tecnologia está adequado e pronto para o uso em produção ou é apenas um protótipo que não pode ser usado em sistemas de produção? Um indicador pode ser a aceitabilidade da comunidade relativamente ao recurso/à ferramenta e o sucesso com que o recurso/a ferramenta são utilizados em sistemas de Tecnologias da Linguagem.

- 0: protótipo preliminar, recurso exemplificativo
- 6: componente imediatamente integrável/utilizável

6 **Sustentabilidade:** Como podem a ferramenta ou o recurso ser mantidos/integrados nos sistemas actuais de Tecnologias da Informação? A ferramenta/o recurso preenchem um certo nível de sustentabilidade no que toca a documentação/manuais, explicação de casos de uso, *frontends*, interface gráfico (GUI), etc.? Utilizam ambientes de programação que seguem boas práticas ou que se encontram standardizados/padronizados (como Java EE)? Existem normas da indústria/investigação e, em caso afirmativo, a ferramenta/o recurso estão de acordo com essas normas (formatos de dados, etc.)?

- 0: acesso restrito, formatos e API ad-hoc
- 6: obedece a normas estabelecidas, totalmente documentado

7 **Adaptabilidade:** Até que ponto os melhores recursos/ferramentas podem ser adaptados/estendidos a novas tarefas, novos domínios/géneros/tipos de texto ou casos de uso?

- 0: praticamente impossível adaptar a ferramenta/o recurso a outra tarefa, impossível mesmo com um grande número de recursos ou pessoas/mês disponíveis
- 6: nível de adaptabilidade muito elevado; a adaptação é muito fácil e eficiente

Tabela de ferramentas e recursos

	Quantidade	Disponibilidade	Qualidade	Cobertura	Maturidade	Sustentabilidade	Adaptabilidade
Tecnologias da Linguagem (Ferramentas, Tecnologias, Aplicações)							
Tokenização, Morfologia (tokenização, anotação morfo-sintáctica, análise/geração morfológica)	4	2	4	5	5	2	5
Análise Sintáctica (superficial ou profunda)	2	4	4	3	4	3	4
Semântica da Frase (desambiguação de acepção, estrutura argumental, papéis semânticos)	1	3	4	3	3	3	4
Semântica do Texto (resolução de co-referência, contexto, pragmática, inferência)	2	1	3	1	2	2	1
Processamento Discursivo Avançado (estrutura textual, coerência, estrutura retórica/RST, identificação de categorias argumentativas (*argumentative zoning)*, argumentação, padrões textuais, tipos de texto, etc.)	1	2	2	2	2	2	2
Recuperação de Informação (indexação de texto, RI multimédia, RI multilingue)	0	0	0	0	0	0	0
Extracção de Informação (reconhecimento de entidades nomeadas, extracção de eventos/relações, reconhecimento de opiniões/sentimentos, mineração de texto)	2	2	4	3	2	1	3
Geração de Linguagem (geração de frases, geração de relatórios, geração de texto)	0	0	0	0	0	0	0
Sumarização, pergunta-resposta, Técnicas avançadas de Acesso à Informação	2	3	4	2	3	1	2
Tradução Automática	3	2	2	3	4	2	2
Reconhecimento de Fala	2	3	4	2	2	2	4
Síntese de Fala	3	3	4	4	4	3	4
Gestão de Diálogo (capacidades de diálogo e modelação do utilizador)	1	1	3	3	4	2	4
Recursos Linguísticos (Recursos, Dados, Bases de Conhecimento)							
Corpora de Referência	4	3	4	5	4	5	5
Corpora: Sintaxe (*treebanks, dependency banks*)	2	3	4	4	4	4	4
Corpora: Semântica	1	1	4	3	3	4	4
Corpora: Discurso	1	1	2	2	1	1	1

	Quantidade	Disponibilidade	Qualidade	Cobertura	Maturidade	Sustentabilidade	Adaptabilidade
Corpora Paralelos, Memórias de Tradução	2	4	3	2	2	3	3
Corpora de Fala (dados em bruto, dados de fala anotados, dados de diálogo)	4	2	4	4	4	3	3
Multimédia e dados multimodais (texto combinado com vídeo/áudio)	0	0	0	0	0	0	0
Modelos de Linguagem	0	0	0	0	0	0	0
Léxicos, Terminologias	5	4	5	4	4	3	3
Gramáticas	1	4	5	2	2	2	2
Dicionário de Sinónimos, WordNets	2	2	4	2	4	3	3
Recursos Ontológicos para Conhecimento do Mundo (e.g. *upper models*, *Linked Data*)	2	2	2	2	3	2	1

Conclusões

A situação do português em relação às Tecnologias da Linguagem tem vindo a melhorar substancialmente, mas requer ainda um esforço contínuo de modo a chegar a um nível de desenvolvimento estável. Devem ser tomadas acções imediatas para que possam ser alcançados progressos importantes para a língua portuguesa.

Existe, para o português, uma série de recursos e ferramentas de processamento, mas muito menos do que para o inglês. Ainda assim, esta comparação tem que ser feita com algum cuidado. Mesmo para o inglês, as tecnologias de apoio à língua que hoje existem encontram-se ainda longe do nível exigido para uma verdadeira sociedade de conhecimento multilingue. De realçar o facto de existir uma rede de centros de investigação, tanto em Portugal como no Brasil, criada para promover o avanço na área das Tecnologias da Linguagem para o português, que deverá assegurar esse desenvolvimento no futuro, se o financiamento não estiver em causa.

Neste relatório foi feita uma primeira tentativa de avaliação do panorama geral de muitas línguas europeias no que diz respeito ao apoio às tecnologias da linguagem, o que permite uma comparação ao mais alto nível e a identificação de lacunas e necessidades.

Para o português, os principais resultados para as tecnologias e os recursos foram os seguintes:

- Apesar de certas sub-áreas específicas deste domínio estarem muito activas, o português é uma língua com poucos recursos, sobretudo se comparada com línguas dos países com maiores investimentos em I&D, como o inglês, o alemão ou o holandês;
- Foram apontados dois grandes *corpora* para o português, sendo que um é pouco representativo, uma vez que apenas abrange

um tipo de texto (jornalístico), e o outro não está totalmente disponível, devido a restrições de direitos de autor;

- Para as variedades do português menos estudadas, têm estado a ser construídos *corpora* nos últimos anos, mas precisam de receber mais atenção;
- Um *corpus* com 1M de palavras anotadas está disponível juntamente com o respectivo etiquetador morfo-sintáctico (*POS tagger*), embora necessite de ser actualizado para estar de acordo com formatos internacionais;
- Em relação às tecnologias da fala, há um conjunto de sistemas comerciais para ambas as variedades europeia e brasileira (reconhecimento de fala, síntese de fala e gestão estatística de diálogo), mas embora as equipas portuguesas e brasileiras sejam muito dinâmicas nesta área, as ferramentas e os *corpora* anotados estão normalmente reservados a uso interno e não estão disponíveis livremente;
- Enquanto muitos *corpora* têm anotação morfo-sintáctica e outros tipos de informação morfológica, os *corpora* com anotação sintáctica são mais raros;
- Foram desenvolvidos alguns analisadores sintácticos, mas eles são, na maioria, ainda muito limitados, tal como os sistemas de sumarização e de resposta-a-perguntas;
- Faltam *corpora* anotados com informação semântica, o que origina uma preocupante situação de falta de ferramentas de processamento e investigação para desambiguação de acepção em português;
- Os corpora paralelos para tradução automática que incluem Português são, sobretudo, os disponibilizados por iniciativas desenvolvidas pela UE e são, consequentemente, muito limitados quanto ao tipo de texto;
- É necessário mais trabalho no que diz respeito a recursos lexicais e wordnets;
- As ferramentas com anotação textual e discursiva são poucas e parciais;
- Quanto mais conhecimento linguístico e semântico uma ferramenta considerar, mais falhas existem (ver, por exemplo, recuperação de informação vs. semântica de um texto);
- É, portanto, necessária mais acção de modo a sustentar um processamento linguístico estável.

Por tudo isto, torna-se clara a necessidade de concentrar mais esforços na criação de recursos para o português, bem como na investigação, na inovação e no desenvolvimento de ferramentas de processamento. A falta de maiores quantidades de dados e a grande complexidade dos sistemas de Tecnologias da Linguagem tornam igualmente indispensável criar novas infra-estruturas para partilha e cooperação.

Sobre a META-NET

A META-NET é uma Rede de Excelência financiada pela Comissão Europeia, actualmente constituída por 47 membros de 31 países europeus. Promove uma Aliança Tecnológica Europeia Multilingue (META), uma comunidade em crescimento que reúne profissionais das Tecnologias da Linguagem e organizações europeias.

Multilingual Europe Technology Alliance (META)

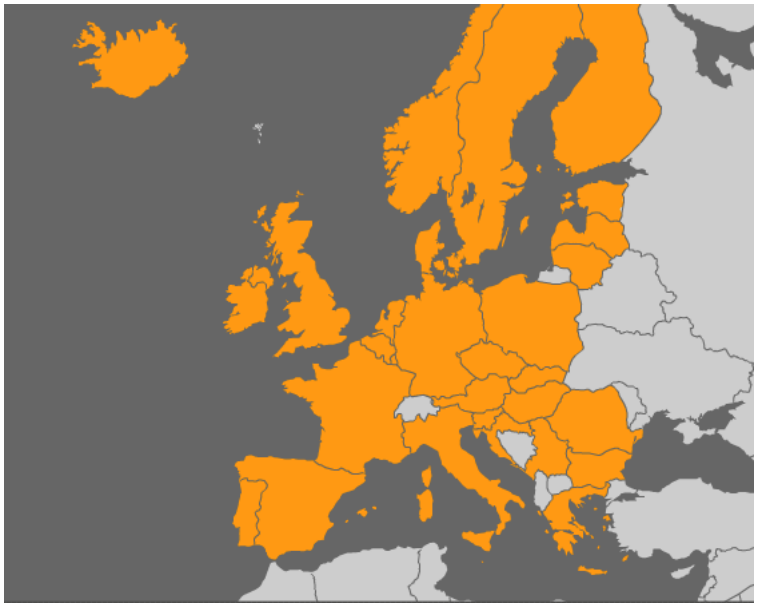


Figura 1: Países representados na META-NET

A META-NET coopera com outras iniciativas, como é o caso do projecto *Common Language Resources and Technology Infrastructure* (CLARIN), que está a contribuir, na Europa, para a investigação na área das Humanidades recorrendo às tecnologias digitais. A META-NET promove as bases tecnológicas para o estabelecimento e a manutenção de uma sociedade de informação europeia verdadeiramente multilingue, que:

- torna possível a comunicação e a cooperação entre línguas;
- permite igual acesso à informação e ao conhecimento em qualquer idioma;
- oferece aos cidadãos europeus tecnologias de informação avançadas e economicamente acessíveis em rede.

A META-NET estimula e promove tecnologias multilingues para todas as línguas europeias. Tecnologias que permitam tradução automática, produção de conteúdos, processamento de informação e gestão de conhecimento para uma grande variedade de aplicações e domínios. Trata-se de uma rede que quer melhorar as actuais abordagens, para que possa ter lugar uma melhor comunicação e cooperação entre línguas. Os europeus têm, independentemente da língua, direitos iguais no acesso à informação e ao conhecimento.

Linhas de Acção

A METANET foi lançada a 1 de Fevereiro de 2010 com o objectivo de alargar a investigação em Tecnologias da Linguagem. A rede estende-se pela Europa, que se une enquanto um mercado digital e um espaço de comunicação único. A METANET tem realizado diversas actividades que promovem os seus objectivos, tendo três

linhas de acção: a META-VISION, a META-SHARE e a META-RESEARCH.

Figura 3: Três linhas de acção na META-NET

A **META-VISION** promove uma comunidade de parceiros dinâmica e influente, que se junta em torno de uma visão e de uma Agenda de Investigação Estratégica comuns.

O principal objectivo desta actividade é a construção, na Europa, de uma comunidade coesa e coerente de Tecnologias da Linguagem, juntando representantes de diversos grupos dispersos de parceiros. No primeiro ano do META-NET houve apresentações, com vista à sensibilização do público, no FLaReNet Forum (Espanha), no Language Technology Days (Luxemburgo), no JIAMCATT 2010 (Luxemburgo), no LREC 2010 (Malta), no EAMT 2010 (França) e no ICT 2010 (Bélgica). De acordo com as primeiras estimativas, o META-NET já contactou com mais de 2500 profissionais das Tecnologias da Linguagem, para com eles levar a cabo estes objectivos e ideias. No META-FORUM 2010, realizado em Bruxelas, os resultados iniciais do desencadear deste processo foram comunicados a mais de 250 participantes, tendo estes, num conjunto de sessões interactivas, mostrado o seu *feedback* relativamente aos dados apresentados pela rede.

A **META-SHARE** criou um sistema aberto, de fácil distribuição, troca e partilha de recursos. A rede *peer-to-peer* de repositórios terá dados linguísticos, ferramentas e serviços web documentados com metadados de alta qualidade e organizados em categorias standardizadas. Os recursos podem ser facilmente acedidos e pesquisados de maneira uniforme. Estes recursos disponibilizados incluem materiais livres, *open source* e restritos, disponíveis comercialmente, mediante uma taxa. A META-SHARE centra-se em dados linguísticos existentes, ferramentas e sistemas, bem como em produtos novos e emergentes necessários à construção e à avaliação de novas tecnologias, novos produtos e serviços. A reutilização, a combinação, a reorientação e a reengenharia de dados linguísticos e ferramentas desempenha um papel crucial. O META- SHARE tornar-se-á, inevitavelmente, uma parte fulcral no mercado das Tecnologias da Linguagem para programadores, especialistas em localização linguística, investigadores, tradutores e profissionais da linguagem de pequenas, médias e grandes empresas. A META-SHARE aborda o ciclo de desenvolvimento completo das Tecnologias da Linguagem – desde a investigação até aos produtos e serviços inovadores. Um aspecto nevrálgico deste trabalho é estabelecer a META-SHARE como uma parte importante e valiosa de uma infra-estrutura europeia global para a comunidade das Tecnologias da Linguagem.

A **META-RESEARCH** constrói pontes para áreas da tecnologia relacionadas. Este projecto visa impulsionar avanços em outras áreas e beneficiar com pesquisas inovadoras que podem favorecer

as Tecnologias da Linguagem. Concretamente, pretende-se trazer mais semântica à tradução automática, optimizar a divisão de trabalho em tradução automática híbrida, explorar o contexto e preparar uma base empírica para tradução automática. A META-RESEARCH colabora com outras áreas e disciplinas, tais como aprendizagem automática e a comunidade da Web Semântica. A META-RESEARCH visa reunir dados, preparar conjuntos de dados e organizar recursos linguísticos para fins de avaliação; compilar inventários de ferramentas e métodos; organizar workshops e tutoriais para membros da comunidade. Este trabalho já identificou, claramente, aspectos da tradução automática em que a semântica pode melhorar as práticas correntes. Paralelamente, fizeram-se sugestões sobre como abordar o problema da integração de informação semântica em tradução automática. A META-RESEARCH está também a terminar um novo recurso linguístico para tradução automática, o *Annotated Hybrid Sample MT Corpus*, que disponibiliza dados para os pares de línguas Inglês-Alemão, Inglês-Espanhol e Inglês-Checo. A META-RESEARCH desenvolveu ainda *software* que reúne *corpora* multilingues que estão espalhados na internet.

Instituições-membros

A tabela em baixo apresenta uma lista das instituições (e seus representantes) que fazem parte da META-NET.

País	Instituição	Representante(s)
Áustria	Universidade de Viena	Gerhard Budin
Bélgica	Universidade de Antuérpia	Walter Daelemans
	Universidade de Leuven	Dirk van Compernolle
Bulgária	Academia das Ciências da Bulgária	Svetla Koeva
Croácia	Universidade de Zagrebe	Marko Tadić
Chipre	Universidade do Chipre	Jack Burston
República Checa	Universidade Charles, em Praga	Jan Hajic
Dinamarca	Universidade de Copenhaga	Bolette Sandford Pedersen and Bente Maegaard
Estónia	Universidade de Tartu	Tiit Roosmaa
Finlândia	Universidade de Aalto	Timo Honkela
	Universidade de Helsínquia	Kimmo Koskenniemi and Krister Linden
França	CNRS/LIMSI	Joseph Mariani
	Evaluations and Language Resources Distribution Agency	Khalid Choukri
Alemanha	DFKI	Hans Uszkoreit and Georg Rehm
	RWTH Aachen University	Hermann Ney

País	Instituição	Representante(s)
	Universidade de Saarland	Manfred Pinkal
Grécia	Instituto para a Linguagem e o Processamento da Fala, "Athena" R.C.	Stelios Piperidis
Hungria	Academia das Ciências da Hungria	Tamás Váradi
	Universidade da Tecnologia e Economia de Budapeste	Géza Németh and Gábor Olaszy
Islândia	Universidade da Islândia	Eirikur Rögnvaldsson
Irlanda	Universidade de Dublin City	Josef van Genabith
Itália	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli"	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Letónia	Tilde	Andrejs Vasiljevs
	Instituto de Matemática e Ciências Computacionais, Universidade da Letónia	Inguna Skadina
Lituânia	Instituto da Língua Lituana	Jolanta Zabarskaitė
Luxemburgo	Arax Ltd.	Vartkes Goetcherian
Malta	Universidade de Malta	Mike Rosner
Holanda	Universidade de Utrecht	Jan Odijk
	Universidade de Groningen	Gertjan van Noord
Noruega	Universidade de Bergen	Koenraad De Smedt
Polónia	Academia Polaca das Ciências	Adam Przepiórkowski and Maciej Ogrodniczuk
	Universidade de Lodz	Barbara Lewandowska-Tomaszczyk and Piotr Pęzik
Portugal	Universidade de Lisboa	Antonio Branco
	Instituto de Engenharia de Sistemas e Computadores	Isabel Trancoso
Roménia	Academia das Ciências Romena	Dan Tufis
	Universidade de Alexandru Ioan Cuza	Dan Cristea
Sérvia	Universidade de Belgrado	Dusko Vitas, Cvetana Krstev and Ivan Obradovic
	Instituto Mihailo Pupin	Sanja Vranes

País	Instituição	Representante(s)
Eslováquia	Academia das Ciências Eslovaca	Radovan Garabik
Eslovénia	Instituto Jozef Stefan	Marko Grobelnik
Espanha	Barcelona Media	Toni Badia
	Universidade Técnica da Catalunha	Asunción Moreno
	Universidade de Pompeu Fabra	Núria Bel
Suécia	Universidade de Gotemburgo	Lars Borin
Reino Unido	Universidade de Manchester	Sophia Ananiadou
	Universidade de Edimburgo	Steve Renals

Referências

[i] European Commission Directorate-General Information Society and Media, *User language*

preferences online, Flash Eurobarometer #313, 2011

(http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).

[ii] European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels,

2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

[iii] UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[iv] European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

[v] Cf. http://observatorio-lp.sapo.pt/pt/dados-estatisticos/falantes-de-portugues) and www.ethnologue.com.

[vi] http://stats.oecd.org/Index.aspx?datasetcode=MIG; http://www.observatorioemigracao.secomunidades.pt/np4/home.html)

[vii] Cf. http://www.portugal-linha.pt/

[viii] Cf. http://cvc.instituto-camoes.pt/index.php; d'Andrade et. al. (1999) orgs. *Crioulos de Base Portuguesa*. Lisboa: APL.

[ix] Cf. Lindley Cintra, L. F. (1971) "Nova proposta de classificação dos dialectos galego-portugueses",

Boletim de Filologia. Lisboa: Centro de Estudos Filológicos, 22, pp. 81-116.

[x]De acordo com os Censos 2001.

[xi] http://www.instituto-camoes.pt/missao-do-instituto-camoes/instituto-camoess-mission.html

[xii] Cf. http://www.internetworldstats.com/stats7.htm.

[xiii] Cf.http://twtrcon.com/2010/02/25/top-5-language-on-twitter-are-english-japanese-portuguese-malay-and-spanish/.

[xiv] Cf. http://www.internetworldstats.com/top20.htm

[xv] Cf. http://mybroadband.co.za/news/internet/15031-Internet-access-Brazil-booms.html.

[xvi] Cf. http://www.internetworldstats.com/stats4.htm, http://www.internetworldstats.com/stats15.htm.

[xvii] Cf. http://www.pordata.pt.

[xviii] http://www.umic.pt/index.php?option=com_content&task=view&id=2777&Itemid=40

[xix] The higher the score, the better the translation, a human translator would get around80.K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of ACL, Philadelphia, PA.