Título dado pela organização do Encontro: Perspectivas de investigação e desenvolvimento no domínio do tratamento computacional da língua e sua importância para a aprendizagem e difusão

Título original: Processamento computacional do Português escrito e falado

Isabel Trancoso
Prof. Catedrática do Instituto Superior Técnico e Investigadora do INESC-ID Lisboa

1. Processamento Computacional da Língua Portuguesa

O Processamento Computacional da Língua Portuguesa tem sido identificado, desde o final da década de 90, como área científica independente pela Fundação para a Ciência e Tecnologia. Este reconhecimento, que o torna independente de áreas de engenharia, linguística, ou literatura onde no passado o poderíamos encontrar, atesta a forte interdisciplinaridade da área, onde se conjugam conhecimentos de processamento de língua natural, processamento de sinal, linguística clássica, psicologia, fisiologia, aprendizagem automática, etc.

Na minha opinião, este foi o passo mais importante dado pelos governos deste país no sentido de promover a utilização do Português como língua de acesso à tecnologia, e vem na sequência de medidas semelhantes tomadas a nível da União Europeia, ao reconhecer na multiplicidade linguística de Europa um património cultural que importa preservar e promover do ponto de vista tecnológico, tendo levado ao estabelecimento de programas na área designada por HLT (*Human Language Technologies*).

Um outro passo muito significativo foi o estabelecimento do projecto Linguateca, um centro de recursos distribuído para o processamento computacional da língua Portuguesa, com três objectivos essenciais: informação, recursos e avaliação.

2. As Várias Vertentes das Tecnologias da Língua

A enumeração exaustiva de todas as áreas que podemos encontrar dentro das chamadas tecnologias da língua seria uma tarefa longa e fora do contexto do presente Encontro, mas conviria distinguir três grandes vertentes: a do processamento da língua natural, a do processamento da fala e a que conjuga ambas e que, à falta de melhor designação, daremos o nome de processamento da língua falada.

Na primeira, que se tem tradicionalmente debruçado sobre o tratamento de texto, inclui-se a análise computacional a variadíssimos níveis: fonológico, morfológico, sintáctico, semântico, pragmático, de discurso, de estilo, etc. Incluem-se também áreas como a geração de texto, a recuperação de informação, a extracção automática de termos, a sumarização, a resposta automática a perguntas, a tradução automática ou assistida e toda a área genérica de interfaces em língua natural.

Dentro da segunda grande área, a do processamento da fala, incluem-se todos os tópicos que lidam com o sinal acústico, como sejam os modelos de produção e percepção de fala, a análise de fala, a codificação de fala para efeitos de transmissão ou armazenamento, a melhoria do sinal de fala, sobretudo em ambientes ruidosos, a síntese de fala ou conversão texto-fala, o oposto ou seja a conversão fala-texto ou reconhecimento de fala, e também o reconhecimento do orador e da língua / dialecto utilizados.

Fazer a ponte entre a fala e o significado subjacente é algo muitíssimo mais complexo e que conjuga as duas grandes áreas anteriores, permitindo aplicar a documentos falados cuja

disponibilidade cresce exponencialmente com as capacidades dos dispositivos de armazenamento existentes hoje em dia e a sua divulgação via internet, todas as técnicas anteriores: resposta a perguntas, classificação semântica, sumarização, extracção automática de termos, recuperação de informação, etc. A compreensão da fala vai muito mais além da simples conversão para texto, envolvendo informação não-linguística e paralinguística e implicando o tratamento das disfluências que caracterizam a fala espontânea. Mais do que construir sistemas de diálogo falados, ambiciona-se construir sistemas verdadeiramente multimodais.
Entre as aplicações do processamento da língua falada de maior interesse para o presente Encontro salientam-se as de âmbito multilingue, em particular o ensino da língua e a tradução fala-para-fala. Talvez um dia, ter um tradutor Português-Mandarim que fale a segunda língua simulando as características vocais do orador original se possa tornar realidade!

3. Aplicações Emergentes

O desenvolvimento de todas estas tecnologias para o Português implicou um grande investimento a nível de construção de recursos linguísticos e de ferramentas básicas. Abre-se agora um enorme potencial de aplicações em áreas emergentes. As tecnologias da língua têm um papel crescente não só em tornar a comunicação pessoa-máquina mais natural, mas também como facilitadoras da comunicação entre humanos. É nesta última área que se inclui, por exemplo, a transcrição/sumarização/indexação de notícias, reuniões, aulas, palestras, tribunais, etc. e a própria tradução automática fala-para-fala.

4. Aplicações para Cidadãos com Necessidades Especiais

Uma das primeiras áreas de aplicação das tecnologias da língua foi a das ajudas para pessoas portadoras de deficiência. A utilização de sintetizadores de fala por invisuais foi um dos grandes motores do seu desenvolvimento, mas há ainda um longo caminho a percorrer no sentido de tornar a fala sintética expressiva e de automatizar a construção de novas vozes. A legendagem automática é também de importância fulcral para portadores de deficiência auditiva. Também os aceleradores de escrita são particularmente importantes para deficiência motora. O potencial de aplicações em reabilitação é enorme, justificando por si só, todo o esforço no sentido de desenvolver e melhorar estas tecnologias.

5. Aplicações ao Ensino da Língua

Pela sua importância no presente encontro, a área de aplicação ao ensino da língua merece uma menção especial. Sendo uma área em que pessoalmente ainda não estive envolvida, pouco posso mencionar para além do enorme potencial que as tecnologias da língua têm para oferecer, não só no ensino de uma língua estrangeira, mas também no ensino da escrita e leitura na primeira língua. Saliento a possibilidade de detecção de erros na escrita, tanto a nível ortográfico como gramatical e de estilo, e também a detecção de pronúncia e entoação incorrectas. O próprio treino da leitura poderia ser facilitado com a interacção com **livros falados digitais**, em que áudio e texto podem estar alinhados palavra a palavra ou frase a frase. Nesta área que é internacionalmente designada como CALL (Computer Aided Language Learning), as tecnologias da língua abrem as portas a novos paradigmas de ensino.
Não gostaria de terminar este ponto sem mencionar um projecto recentemente finalizado, cuja responsável, a Profª Maria Helena Mira Mateus, por motivos de força maior não pode comparecer neste Encontro. O projecto “Diversidade Linguística na Escola Portuguesa”, levado a cabo pelo ILTEC em colaboração com o Ministério da Educação e apoio da Fundação Calouste Gulbenkian, recolheu produções orais e escritas de crianças que têm como língua materna o Crioulo de Cabo

Verde, o Guzerate, o Mandarim e o Ucraniano. Nele foram compiladas informações de natureza linguística e sociolinguística e proposta uma tipologia de modelos de ensino para a diversidade linguística.

6. Processamento de Outras Variedades do Português

O sexto lugar que o Português ocupa na lista das línguas mais faladas do mundo traz-nos responsabilidades significativas, sendo uma delas o desenvolvimento de tecnologias da língua para o Português tendo em conta a sua portabilidade para outras variantes. Isto implica um esforço muito grande na construção de recursos linguísticos que cubram todas as variantes, algo que ainda só existe para o Português falado em Portugal e no Brasil.

7. Panorama de I&D em Tecnologias da Língua em Portugal

Não queria terminar esta breve apresentação sem mencionar os principais intervenientes a nível das tecnologias da língua em Portugal. Existem centros de investigação ligados às Universidades de Lisboa, Porto, Coimbra, Aveiro, Minho, Algarve, Évora e Beira Interior, mencionando apenas os mais activos. Existem também centros de I&D em empresas operadoras da rede telefónica (Vodafone, PT Inovação), editoras (Porto Editora), produtoras de software (Priberam, Microsoft), etc.

Estes intervenientes têm tido um papel mais ou menos activo na avaliação conjunta de ferramentas promovida pelo projecto Linguateca e que envolveu já o reconhecimento de entidades mencionadas para o Português e a recolha de informação cruzada e resposta automática a perguntas. Na última, realizada a nível internacional, há que realçar o 1º prémio conquistado pela empresa Priberam.

8. Conclusões e Demonstração

Nesta apresentação, tentou-se salientar o impacto das tecnologias da língua, citando algumas das suas principais vertentes e aplicações. Há no entanto ainda muito que fazer na promoção da utilização das interfaces em língua natural, na mobilização das empresas para a integração de tecnologias da língua nos seus produtos e na dinamização da utilização do Português como língua de acesso à tecnologia em todas as suas variantes.

Como demonstração do estado da arte em algumas das tecnologias mencionadas, gostaria de ilustrar a legendagem automática de notícias (aplicada ao telejornal de dia 8 de Dezembro) e a síntese áudio-visual da frase que dirijo sobretudo aos participantes chineses: sejam bem vindos a Portugal.