Os recursos da Linguateca ao serviço do desenvolvimento da tecnologia de fala na Microsoft

Daniela Braga, Miguel Sales Dias

i-dbraga@microsoft.com; midias@microsoft.com

Microsoft Language Develpment Center, Porto Salvo, Portugal

Apesar de o português ser uma das línguas mais faladas do mundo enquanto língua materna (com cerca de 235 milhões de falantes) e língua oficial de 9 estados independentes (Angola, Brasil, Cabo Verde, Guiné Bissau, Moçambique, Portugal, São Tomé e Príncipe, Timor), se recuarmos uma década apenas, era clara a escassez de recursos disponíveis para as comunidades científicas da Linguística e da Engenharia da Linguagem, sobretudo em formatos inteligíveis para processamento computacional. A Linguateca veio preencher com sucesso essa lacuna, contribuindo não só para a aproximação entre as comunidades científicas portuguesa e brasileira que trabalham em Processamento da Linguagem Natural, Linguística Computacional e áreas relacionadas, como também para a divulgação de trabalhos académicos e para a disponibilização livre de recursos linguísticos (corpora), metodologias de avaliação, ferramentas computacionais e resultados de projectos de I&D, nestes domínios, que de outra forma se manteriam dispersos e dificilmente acessíveis.

Nesta comunicação, é nosso objectivo mostrar o papel fundamental que os recursos linguísticos em Português disponibilizados pela Linguateca (com especial destaque para o CETEMPúblico, CETENFolha e COMPARA), têm desempenhado no desenvolvimento, teste e avaliação dos algoritmos de processamento da linguagem natural que integram os sistemas de síntese de fala em português europeu e em português do Brasil, actualmente em desenvolvimento no Microsoft Language Development Center – MLDC, para os produtos do Exchange 14 e Office Communication Server 13, que sairão no mercado português em 2009.