Capítulo 5

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

Daniela Braga e Miguel Sales Dias

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 5, p. 29–33, 2008

No contexto das comunidades científicas do processamento da linguagem natural, linguística computacional e áreas relacionadas, como o processamento da fala, é consensual dizer que o panorama do processamento computacional do português não seria definitivamente o mesmo, sem a existência do projecto e do grupo de interesse dinamizado pela Linguateca. Com efeito, 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 8 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 afins, como também para a divulgação de trabalhos académicos e para a disponibilização livre de recursos linguísticos, metodologias de avaliação, ferramentas computacionais e resultados de projectos de I&D, nestes domínios, que de outra forma se manteriam dispersos e dificilmente acessíveis.

Assim se compreende a surpresa e alguma consternação com que a comunidade científica, recebeu a notícia de que o projecto da Linguateca iria terminar no final do ano civil de 2008, por exaustão do respectivo financiamento público. A Linguateca tem desempenhado ao longo dos seus 10 anos de existência um papel catalisador de sinergias oriundas das várias comunidades portuguesa e brasileira que trabalham no processamento computacional do português e teve como resultado a produção de valiosos recursos linguísticos, nomeadamente, de texto processados e de motores de utilização proveitosa dos mesmos, todos eles com direitos cedidos ao domínio público e assim disponibilizado gratuitamente à comunidade científica. Para além desta, outras mais-valias saídas do projecto Linguateca, são de realçar, nomeadamente, a criação de uma comunidade científica especializada na produção de recursos linguísticos e na produção de trabalho de valor académico ao mais alto nível, a organização das campanhas mais estendidas de avaliação de recursos em português – como as Morfolimpíadas (Santos e Costa, 2003) e o HAREM (Santos e Cardoso, 2007) – ou, ainda, o apoio à participação dos recursos para o português no CLEF (Cross-Language Evaluation Forum, Forum de avaliação entre várias línguas). De destacar ainda, o trabalho tão útil à comunidade, de repositório de publicações científicas relacionadas com o processamento computacional do português que o sítio da Linguateca tem proporcionado, permitindo consultar com facilidade trabalhos de referência na área, teses de mestrado e doutoramento, etc, sendo ainda acessível através de um sistema de busca assistida, o SUPeRB (Cabral et al., 2008).

5.1 A experiência da indústria: o impacto da Linguateca no desenvolvimento de produtos na Microsoft

Mais do que destacar as virtualidades sobejamente conhecidas trazidas pela comunidade que criou e desenvolveu o universo Linguateca, gostaríamos de deixar o nosso testemunho enquanto membros integrantes da indústria, a qual entende que as sinergias estratégicas academia-indústria são factores que propiciam a inovação e a melhoria da qualidade do desenvolvimento de produtos de software.

O MLDC – Microsoft Language Development Center, sediado no Tagus Park, Porto Salvo, é um grupo de produto da Microsoft onde os autores trabalham, integrado num ambiente de desenvolvimento distribuído, que inclui, para além de Portugal, pólos em Redmond/EUA e Pequim/China. Este grupo alargado produz as tecnologias de reconhecimento e síntese de fala, utilizadas pelos diversos grupos de produto da companhia nos seus desenvolvimentos de software, tais como os grupos Live, Servidor, Cliente (Windows), Mobilidade, Automóvel e Entretenimento. O MLDC participa em todas as acividades centrais do grupo de fala, tais como a expansão das tecnologias de fala para um número elevado de línguas, onde se inclui o desenvolvimento da síntese de fala em português europeu e português do Brasil. Muito naturalmente o MLDC ficou incumbido de desenvolver as tecnologias de fala para estas duas variantes do português. Este artigo é, assim, uma oportunidade para os autores relatarem o impacto extremamente positivo dos recursos linguísticos disponibilizados pela Linguateca, no desenvolvimento de tecnologias de síntese de fala, no MLDC.

De facto, conhecedores do projecto e sítio da Linguateca, os engenheiros e linguistas do MLDC cedo começaram a utilizar de forma assídua alguns dos recursos disponíveis, para o 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. Podemos salientar a utilização dos seguintes recursos: CETEMPúblico (Rocha e Santos, 2000), CETENFolha, COMPARA (Frankenberg-Garcia e Santos, 2003) e a Floresta Sintá(c)tica (Bosque e Floresta Virgem) (Afonso et al., 2001). Todos estes recursos, com excepção do COMPARA (corpus paralelo em português e inglês nos dois sentidos), podem ser descarregados mediante o preenchimento de um formulário simples em `www.linguateca.pt`. Foram várias as tarefas executadas sobre os corpora CETEMPúblico, CETENFolha e COMPARA, de forma a torná-los adequados ao desenvolvimento de produto na Microsoft:

1. Selecção automática de frases a serem gravadas para a construção de uma base de dados de talentos de voz;
2. Selecção de casos de teste, saídos de corpora de texto real, para validação de algoritmos de:
   a. Separação de frases;

   b. Separação de palavras;
   c. Normalização de texto;
   d. Desambiguação de homógrafos;
   e. Conversão grafema-fone;

3. Obtenção e generalização de padrões para criação de regras de normalização de texto;

4. Obtenção de listas de frequência de léxico em certos domínios.

Os corpora anotados morfológica e sintacticamente, como a Floresta Sintá(c)tica, tiveram uma outra utilidade para nós. Quer o Bosque (inclui os primeiros 1000 excertos do CETEMPúblico e do CETENFolha revistos por linguistas), quer a Floresta Virgem (é composta pelo primeiro milhão de palavras do CETEMPúblico e do CETENFolha anotado automaticamente pelo analisador sintáctico PALAVRAS (Bick, 2000)), foram usados como corpus de treino do nosso analisador sintáctico automático, cujo resultado é depois utilizado pelo módulo de desambiguação de homógrafos e cuja desambiguação é também obtida através da análise do contexto sintáctico.

Gostaríamos de salientar ainda a escassez de recursos de texto e corpora anotados morfossintacticamente de larga dimensão existentes para o português, para além daqueles que são disponibilizados pela Linguateca. Na verdade, um exercício simples de busca no catálogo da ELRA (European Language Resources Association, `catalog.elra.info`) ou da APPEN (`www.appen.com.au`), dois catálogos de recursos linguísticos de referência na indústria, mostra que não existem muitos mais recursos disponíveis no mercado para o português com qualidade e quantidade significativas, pelo que o desenvolvimento dos nossos sintetizadores nas duas variedades do português se baseou quase exclusivamente nos recursos disponibilizados pela Linguateca.

5.2 Conclusão

A cessação do financiamento público da Linguateca não deve ser encarada como o fim deste projecto, ou das comunidades que ajudou a sensibilizar e a dinamizar. Antes pelo contrário, defendemos que deve ser entendido como uma hipótese de renovação, de regeneração e renascimento. Vivemos numa sociedade de informação e comunicação em rápido processo de maturação e transformação, que necessita de sistemas de busca eficientes, de sistemas de tradução automática para se mover na babel linguística, de tecnologias de fala para facilitar a interacção com as máquinas, só para citar alguns exemplos. Toda esta panóplia tecnológica de texto e de fala assenta em corpora mais extensos, de géneros mais diversificados, com processos de busca mais eficientes. A Microsoft necessitará sempre de recursos linguísticos com qualidade (corpora de texto, corpora paralelos, léxicos fonéticos,

corpora de fala transcrito e anotado), para melhorar os seus sistemas de síntese (Braga et al., 2008) e reconhecimento de fala, os seus sistemas de busca (MSN, `www.msn.com`) e os seus tradutores automáticos (Translator, `www.windowslivetranslator.com`). Por outro lado, a Microsoft verifica que continuam a existir lacunas na oferta de recursos para o português. As actividades desencadeadas e dinamizadas pela Linguateca permitiram formar pessoas capazes de suprir essas lacunas, pessoas essas que são hoje especialistas na produção de recursos linguísticos para o português e que podem inclusive expandir esse conhecimento para a outras línguas. A indústria e a sociedade esperam assim que o ecosistema de processamento computacional do português, dinamizado pela Linguateca, se mantenha vivo e activo.