A Linguateca e o projecto "Processamento computacional do português"

Cet article décrit Linguateca, centre de ressources pour le traitement automatique de la langue portugaise, qui a commencé son activité en 2000, à la suite du projet "Processamento computacional do português" qui s'est déroulé entre 1998 et 2000. Après une brève introduction historique, il présente une description du travail réalisé, encadré dans les trois principaux vecteurs dans lesquels s'oriente Linguateca : dissémination d'informations, développement de ressources et promotion d'évaluations communes.

Este artículo describe Linguateca, un centro de recursos para el procesamiento computacional de la lengua portuguesa, que empezó su actividad en el año 2000, continuando el proyecto "Processamento computacional do português" desarrollado entre 1998 y 2000. Tras una breve introducción histórica, explicando algunos de los presupuestos básicos de Linguateca, se presenta una descripción del trabajo realizado, encuadrado dentro de los tres pilares que definen Linguateca: diseminación de información, desarrollo de recursos y promoción de evaluaciones conjuntas.

Introdução

A Linguateca [7] [19] [21] [37] é um centro de recursos para o processamento computacional da língua portuguesa, em todas as variantes. Neste artigo apresentamos o trabalho já feito, a nossa actividade presente e as nossa expectativas de futuro.

Começamos por fornecer alguma informação de contexto: A Linguateca é a continuação natural do projecto "Processamento computacional do português" [17], que decorreu no SINTEF, em Oslo, de Maio de 1998 a Maio de 2000, dado que a área da engenharia da linguagem do português, considerada prioritária pelo Ministério da Ciência e da Tecnologia (MCT) português da altura [26], sofria de uma patente falta de planeamento. Nesse sentido, este projecto surgiu como a forma de o MCT promover o planeamento e reforçar o processamento da língua portuguesa, dando-lhe lugar no Livro Branco em Ciência e Tecnologia [25] e nos debates públicos sobre política científica que o precederam [18].

A intervenção da Linguateca baseia-se no modelo IRA – Informação, Recursos e Avaliação. De facto, foi por esta ordem que as actividades da Linguateca (e do projecto que a precedeu) foram evoluindo: primeiro, fez-se um levantamento do que existia e criou-se uma forma de divulgar essa informação, o portal www.linguateca.pt; depois, partiu-se para a criação dos recursos considerados mais prioritários, levando em consideração o referido levantamento; finalmente, o trabalho voltou-se para a avaliação dos recursos já existentes na comunidade, não só para tornar possível medir a evolução no futuro, como para obter indicadores sobre o grau de maturidade de cada sub-área.

Ao longo do tempo, a Linguateca foi estabelecendo cooperação com diversas instituições, nalguns casos criando pólos no interior das mesmas. O objectivo destes pólos é tornar o trabalho efectuado nestas instituições mais visível, fornecendo a infraestrutura de forma a partilhar os resultados com a comunidade em geral.

Foi nesse sentido que a Linguateca se tornou numa organização distribuída, abrindo pólos 1) em Braga, no Departamento de Informática da Universidade do Minho, 2) em Lisboa, no LabEL, Centro de Automática da Universidade Técnica de Lisboa do Instituto Superior Técnico, 3) em Odense, no projecto VISL, na Universidade da Dinamarca do Sul), 4) em Lisboa, dando origem ao projecto COMPARA, 5) no Porto, no Centro de Linguística da Faculdade de Letras da Universidade do Porto, e ainda 6) em Lisboa no grupo XLDB da Faculdade de Ciências da Universidade de Lisboa, e tendo colaboração estreita com vários centros brasileiros de linguística computacional, como por exemplo o NILC[1].

Há alguns pressupostos fundamentais da Linguateca que convém realçar: em primeiro lugar, a nossa matéria prima é a língua portuguesa, independentemente da variante, e por isso tentamos produzir serviços e recursos abrangendo todas as comunidades que falam português. Em segundo lugar, a nossa actividade destina-se a servir a comunidade com um todo, e daí a insistência na criação de recursos publicamente disponíveis, quer para universidades e centros de investigação, quer para empresas. Finalmente, mantemos que é essencial que sejam os falantes do português a tomarem em ombros a tarefa de avançar a área, em vez de se subordinarem ao modelo do inglês [20].

Informação

Um dos requisitos mais importantes para um bom planeamento consiste em saber qual o trabalho já feito. Foi nesse sentido que o projecto começou por fazer um levantamento do que já tinha sido feito na área da engenharia da linguagem do português, bem como dos recursos que poderiam ser utilizados por quem trabalhe nessa área [39].

Este levantamento permitiu estabelecer prioridades, e decidir o que era mais premente para o progresso da área. Permitiu também chegar à conclusão de que havia grande duplicação de esforços: diferentes grupos trabalhavam nos mesmos assuntos sem conhecerem o trabalho uns dos outros; havia uma falta de divulgação notória do trabalho efectuado na área. Construímos assim um portal na rede (Internet) tentando cobrir toda a actividade na área, além de fornecer informações várias de utilidade ao trabalho no processamento do português.

A manutenção deste portal é uma das tarefas que a Linguateca tem desempenhado desde a sua criação, produzindo um catálogo actualizado de recursos, ferramentas, actores, publicações e outras informações interessantes na esfera do processamento do português. Mantemos também um fórum onde divulgamos notícias, oportunidades de emprego, conferências e cursos na área, assim como a equipa da Linguateca responde a sugestões e comentários, esclarece dúvidas e tenta dar apoio a todos os membros da comunidade.

Na tabela 1 podemos ver que o interesse pelo nosso portal tem vindo a crescer como aliás o conteúdo servido pelo mesmo. A quebra observável no ano de 2003 foi provocada, tudo leva a crer, pela mudança de endereço (URL) no final de 2002.

Ano	Nº de visitas ao portal
2004 (11 primeiros meses)	657.130
2003	315.917
2002	406.298
2001	239.224
2000	124.966
1999	60.658
1998 (segundo semestre)	3.185
Total	**1.807.378**

Tabela 1: Visitas ao nosso portal

Na tabela 2 verifica-se, como seria natural, que as visitas para as quais conseguimos determinar a origem são, na sua grande maioria, efectuadas do Brasil ou de Portugal, onde se encontram maioritariamente as pessoas que trabalham em engenharia da linguagem em português.

Origem	Nº de visitas ao portal
Brasil	426.062
Portugal	211.919
Espanha	16.524
Grã-Bretanha	12.986
Alemanha	11.085
Outros países europeus	56.903
Estados Unidos	17.667
América (sem Brasil e Estados Unidos)	18.485
África, Ásia e Oceânia	13.906
Indeterminado	1.021.841
Total	1.807.378

Tabela 2

Recursos

AC/DC

O serviço AC/DC [10] [12] é o resultado da constatação de que fazia falta uma forma expedita de aceder aos corpora existentes, mesmo que fossem de livre acesso. Além de os juntar num único ponto de acesso com uma interface mais uniforme, o que foi um passo em frente, os corpora acessíveis a partir do serviço AC/DC foram ainda enriquecidos com:

a) segmentação em frases, parágrafos e unidades textuais;
b) classificação gramatical e análise sintáctica fornecidas pelo analisador sintáctico PALAVRAS [22] de Eckhard Bick.

Neste momento o AC/DC permite fazer pesquisas em corpora de textos jornalísticos, literários, didácticos e de correio electrónico, num total de mais de 250 milhões de palavras em português.

COMPARA

O inglês é sem dúvida nenhuma o idioma mais traduzido para português. O serviço COMPARA [2] [3], lançado em

colaboração com Ana Frankenberg-Garcia, dá acesso a um corpus paralelo de traduções de português para inglês e vice-versa. Oferece um grande número de opções de procura e uma interface amigável para diferentes tipos de utilizadores com diferentes necessidades de informação contrastiva.

O COMPARA é sem dúvida o maior corpus paralelo editado e revisto contendo o português. Neste momento (versão 6.0), inclui 53 textos originais e 56 traduções, e está em constante crescimento, tendo mais de uma dezena de textos em lista de espera para serem adicionados ao corpus, e tendo sido recentemente iniciada a sua anotação morfossintáctica.

CETEMPúblico e CETENFolha

Embora os projectos AC/DC e COMPARA permitissem aos investigadores estudar a língua quantitativamente e extrair várias informações interessantes, não era possível distribuí-los na íntegra, o que dificultava, por exemplo, o teste de programas sobre o material neles contido. Por outro lado, sabia-se que a existência de corpora de referência de grande dimensão em português seria bastante útil. Os corpus CETEMPúblico e CETENFolha foram uma tentativa de responder a essa lacuna.

O CETEMPúblico (Corpus de Extractos de Textos Electrónicos MCT/Público) [14] [35] contém aproximadamente 180 milhões de palavras em português de Portugal. Foi criado pelo projecto Processamento computacional do português após a assinatura de um protocolo entre o Ministério da Ciência e da Tecnologia (MCT) português e o jornal PÚBLICO em Abril de 2000. O CETENFolha (Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo) contém cerca de 24 milhões de palavras em português brasileiro, com base nos textos do jornal Folha de S. Paulo que integravam o corpus NILC/São Carlos, compilado pelo Núcleo Interinstitucional de Lingüística Computacional (NILC).

Além de serem acessíveis através do AC/DC, estando portanto segmentados em frases, parágrafos e unidades textuais e classificados gramaticalmente e sintacticamente pelo analisador sintáctico PALAVRAS, tanto o CETEMPúblico como o CETENFolha, podem ser obtidos na íntegra de um dos nossos servidores, bastando para isso que o utilizador se registe.

Floresta Sintá(c)tica

Outra das lacunas detectadas ao inventariar a área do processamento do português, foi a inexistência, para a nossa língua, de um "treebank", ou seja, de um conjunto significativo de árvores sintacticamente analisadas que permitem estudos sintácticos e avaliação de analisadores automáticos. Com a Floresta Sintá(c)tica [40] [41], lançada em 1999 em colaboração com Eckhard Bick e o projecto VISL[2], tentámos suprir essa lacuna.

Este projecto de grande envergadura conseguiu já coligir 7.756 árvores, analisadas pelo PALAVRAS e posteriormente revistas por linguistas, correspondendo a aproximadamente 150 mil palavras (versão 6.2).

AnELL

O AnELL [6] foi o resultado da conjugação das vontades, tanto do LabEL[3], como da Linguateca, em fornecer um serviço gratuito de anotação linguística de textos, em casos em que os utilizadores não pudessem ceder esses textos para consulta pública. Este serviço, acessível na rede, utiliza o INTEX [33], um sistema de desenvolvimento de sistemas de processamento de linguagem natural, para produzir a anotação linguística, com base nos recursos linguísticos do LabEL [23].

Oferece dois tipos de anotação: totalmente automática ou semi-automática. Nesta última, em fase de arranque, os resultados da análise automática são (parcialmente) revistos por um linguista.

Corpógrafo

O Corpógrafo [28] [29], desenvolvido no pólo do Porto da Linguateca permite, através de uma interface simples na rede, compilar e pesquisar corpora especializados (muitas vezes pertença exclusiva de um único utilizador) sem exigir a este conhecimentos avançados de informática.

Fornece um ambiente de trabalho que pretende resolver os problemas práticos das pessoas, dos mais básicos aos mais complexos. Por exemplo, por um lado, o Corpógrafo tem ferramentas (independentes) que extraem texto de PDF, por outro, também tenta semiautomaticamente extrair definições. Uma das suas principais funcionalidades é a extracção semi-automática de terminologia, permitindo também criar bases de dados terminológicas.

O Corpógrafo tem um artigo independente nesta edição da revista *Terminometro* [4].

NATools

O NATools é um conjunto de programas desenvolvidos no pólo de Braga da Linguateca para alinhamento – ou seja, interligação de corpora paralelos [1]. O NATools inclui, além de um alinhador de frases e outro de palavras/termos, também um conjunto de ferramentas para trabalhar com corpora alinhados: um gerador de dicionários probabilísticos acessíveis pela rede; um módulo de classificação/avaliação da probabilidade de tradução de dois textos; um extractor de terminologia bilingue