Estudando o português tal como é usado: o serviço AC/DC

Luís Fernando Costa, Diana Santos, Paulo Rocha

Linguateca, Pólo de Oslo, SINTEF ICT, Noruega

{luis.costa,diana.santos}@sintef.no, paulo.rocha@di.uminho.pt

Abstract. *The AC/DC service has been giving access to Portuguese corpora through the Web since 1999. This paper describes the tasks related to processing and making the texts publicly available. It also provides an overview on the interface with which the users can query the corpora and finalizes pointing future directions.*

Resumo. *O AC/DC é um serviço que desde 1999 dá acesso a corpos em português através da Internet. Neste artigo descrevemos sucintamente o processo pelo qual os textos são processados e tornados públicos e a interface através da qual se podem fazer as pesquisas. Concluímos lançando pontes para o desenvolvimento futuro deste serviço.*

1. Introdução

O objectivo principal do serviço AC/DC (http://www.linguateca.pt/ACDC/) consiste em possibilitar o estudo da língua portuguesa tal como ela é usada pelos seus falantes (as pessoas que falam e escrevem português). Os utilizadores do AC/DC podem obter facilmente exemplos reais dos fenómenos linguísticos que pretendem estudar e/ou obter dados quantitativos sobre os mesmos fenómenos.

Os conteúdos disponibilizados pelo AC/DC têm crescido ao longo do tempo, desde 1999. A escolha dos mesmos tem sido motivada, por um lado, pelas oportunidades que foram surgindo (pessoas que disponibilizaram textos, etc.), mas também por tentarmos dar uma imagem global do português (diferentes tipos de textos, diferentes variantes, etc.). A interface do serviço foi reformulada em 2007 precisamente para responder ao facto de os diferentes corpos terem características diversas, o que aconselhava a que a interface se adaptasse e adequasse ao corpo com que um utilizador estivesse a trabalhar.

O trabalho descrito no artigo foi desenvolvido no âmbito da Linguateca, co-financiada pelo governo português, pela União Europeia (FEDER e FSE), sob o contracto POSC/339/1.3/C/NAC, pela UMIC e pela FCCN.

2. Construção dos corpos

Nesta secção descrevemos de forma bastante simplificada o processamento aplicado a todos os corpos do projecto AC/DC.

De forma a integrá-los numa plataforma comum, todos os corpos são pré-processados de forma a terem uma codificação comum da informação estrutural. Esta informação tem duas vertentes distintas: (i) a identificação da estrutura dos textos tanto na sua disposição gráfica (marcação de parágrafos) e estrutura textual (divisão em

capítulos, entrevistas ou notícias, identificação de títulos, de listas, de notas de rodapé, etc.) como na sua identificação linguística (o reconhecimento das frases); (ii) a marcação de alguma informação extratextual associada à origem dos textos: data, género, secção de jornal, autor, variante, etc. ou relacionada com a sua incorporação no corpo: separação em extractos, numeração, etc.

A fase seguinte é a anotação morfo-sintáctica, que é feita com o PALAVRAS [Bick 2000]. O PALAVRAS atribui, a cada unidade do corpo, o seu lema, a sua categoria gramatical e outras características morfo-sintácticas, e a sua função sintáctica. Adicionalmente, o analisador tenta identificar, através de heurísticas morfológicas, palavras não constantes do dicionário.

De seguida, aplica-se um conjunto de programas que transformam o corpo anotado naquilo a que chamamos o "formato AC/DC". Este consiste num conjunto de campos separados por caracteres de tabulação, que, para além da informação extraída do resultado do PALAVRAS, voltam por exemplo a reunir contracções que são desdobradas pelo PALAVRAS, como explicado em [Santos & Bick 2000].

A fase final da criação de um corpo no AC/DC consiste em codificar os textos neste formato com o IMS Corpus Query Processor (CQP) [Christ et al. 1999], ferramenta da qual as funções de pesquisa do AC/DC tiram partido intensivamente.

3. Interface

A interface foi desenhada para permitir aos utilizadores com reduzidos conhecimentos de informática efectuarem pesquisas nos corpos.

3.1. Opções gerais

Para todos os corpos, a interface permite procurar palavras individuais pela sua forma (inclusive usando expressões regulares), ou usando os vários atributos criados a partir da anotação dos corpos pelo PALAVRAS, nomeadamente, lema, categoria gramatical, tempo verbal, caso pronominal, pessoa, número, género, e função sintáctica. É possível igualmente obter a distribuição dos resultados por cada um destes atributos (por exemplo, descobrir quantas vezes num corpo a forma *abandono* é um verbo e quantas um substantivo). Essas distribuições podem ser obtidas por ordem de frequência ou alfabética. Mas certamente que as concordâncias (e as distribuições) não são limitadas a palavras individuais, e a sintaxe do CQP permite procurar expressões arbitrariamente complexas, como ilustrado em [Santos 2008].

Outro serviço do AC/DC com uma interface própria é o Ordenador, que permite consultar a quantidade de ocorrências de determinada forma ou lema em qualquer dos corpos, ou na sua totalidade. Esta interface aproveita as listas de formas e lemas que são criadas automaticamente aquando da criação de cada corpo, permitindo também fazer procuras de palavras relacionadas, como por exemplo todas as palavras iniciadas por "caix", *caix.**.

3.2. Opções individualizadas por corpo

À medida que fomos adicionando mais corpos ao projecto, os tipos de texto foram-se diversificando, e algumas das opções iniciais foram-se revelando inadequadas para os novos corpos, ou não respondendo inteiramente a essa diversidade. Assim, foi criada em

2007 uma nova interface em PHP, que usa ficheiros de configuração distintos para cada corpo, evitando assim apresentar opções de busca desnecessárias. Nesse sentido, mostram-se apenas os atributos que são relevantes para cada um dos corpos: por exemplo, autor, obra e tipo de texto (para os corpos literários), semestre e secção do jornal (para o CETEMPúblico e o CETENFolha), variante (para o CDHAREM, CONDIVport e Museu da Pessoa), década, fonte e tema (CONDIVport), etc.

4. Trabalho futuro

4.1 Estudos

Há certos tipos de estudos que o AC/DC permite e que gostaríamos de poder (nós ou outras pessoas) repetir para o português ou pelos menos confirmar, tais como: as propriedades estatísticas da língua: por exemplo, que itens ou palavras têm um padrão fácil de prever, e quais as que são imprevisíveis ou de ocorrência inesperada [Curran & Osborne, 2002]; ou a afirmação de [Davies 2005, p. 321] de que "para uma língua como o espanhol, (...) há relativamente poucas formas, como "ser", que têm uma frequência elevada com categorias gramaticais diferentes, ou com lemas diferentes, e que não consigam ser facilmente desambiguadas num contexto muito limitado"; questões morfológicas: por exemplo o género dos neologismos, o contexto das palavras sem género intrínseco, o "outro género" de substantivos que se podem referir aos dois sexos, como *girafa* ou *presidente*, ou quando se usa o infinitivo impessoal, ou ainda quando se usam dois adjectivos coordenados, seguidos, ou abraçando o núcleo; questões semânticas: tais como propriedades temporais como a distribuição dos tempos, a distribuição de advérbios de tempo, a menção ao passado e ao futuro, ou propriedades espaciais: que relações e entre que "objectos" linguísticos; ou a expressão da causalidade em português, etc.; questões sintácticas e discursivas: orações relativas, peso dos constituintes, tamanho dos sintagmas, a posição dos clíticos, etc.; e questões associadas ao género textual: havendo vários tipos de textos no AC/DC poder-se-iam estudar indicadores dos vários géneros, ou verificar características já avançadas sobre cada um.

Naturalmente, uma das questões mais interessantes é precisamente a das diferenças entre o português do Brasil e o de Portugal, e nesse aspecto o AC/DC é único ao permitir esse estudo segundo a metodologia avançada por [Silva 2008a,b].

4.2 Funcionalidades

Ao longo dos anos temo-nos deparado com muitas formas de melhorar e estender o número de possibilidades oferecidas pelo AC/DC, muitas delas sugeridas por utilizadores, ou por projectos semelhantes[1]. Alguns exemplos são: permitir a utilizadores a contribuição de listas ou padrões que possam ser usados por eles ou por outros, à semelhança do Corpus do Português (http://www.corpusdoportugues.org/), mas tornando-os públicos; no cálculo de frequências, deixar os utilizadores definirem as classes que pretendem (as faixas); obtenção automática de "bons" exemplos para questões de lexicografia [Kilgarriff et al. 2008]; procura de frases semelhantes, com a

[1] Pode-se consultar uma lista de projectos semelhantes no catálogo de recursos da Linguateca (http://www.linguateca.pt).

consequente escolha (e mesmo definição) de diferentes medidas de semelhança; testes para ensino da redacção em português, como descritos em [Santos 2008]; comparação entre dois itens, à semelhança da possibilidade de [Davies 2005]; obter distribuições cruzadas (mais do que uma categoria simultaneamente); criar automaticamente resultados gráficos para as distribuições; criar um meta-corpo contendo todos os corpos para maior facilidade de comparação entre os géneros textuais.

4.3 Ajuda à melhoria da anotação

Outra coisa que pretendemos mudar no futuro é a possibilidade de mais facilmente revermos a anotação sintáctica dos corpos usando os próprios utilizadores como correctores ou pelo menos como deflagradores de uma revisão. Com efeito, muitas vezes é preciso rever e corrigir, para estudos empíricos, alguma parte da anotação, ou mesmo suplementar com outras questões. Se fosse possível partilhar essa nova anotação com toda a comunidade isso seria uma mais-valia incomparável.

De qualquer maneira, mesmo para permitir o desenvolvimento e melhoria na própria Linguateca seria preciso definir um ambiente colaborativo mais eficiente.

Referências

Eckhard Bick (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Oliver Christ, Bruno M. Schulze, Anja Hofmann e Esther Koenig (1999). The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. University of Stuttgart, 8 de Março de 1999 (CQP V2.2)

James Curran e Miles Osborne (2002). A very very large corpus doesn't always yield reliable estimates. *Joint CoNLL02 - Workshop on Very Large Corpora*, Taipei.

Mark Davies (2005). The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics* 10(3): 301-28.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell e Pavel Rychlý (2008). GDEX: Automatically finding good dictionary examples in a corpus. Em: Proceedings of EURALEX 2008, Barcelona, Espanha.

Diana Santos e Eckhard Bick (2000). Providing Internet access to Portuguese corpora: the AC/DC project. Em Maria Gavrilidou et al, editores, Proceedings of LREC 2000. (Atenas, Grécia, 31 de Maio a 2 de Junho de 2000), páginas 205-210.

Diana Santos (2008). Corpos linguísticos da Linguateca: apresentação, TaLC at TaLC: Teaching and Linguateca's (Portuguese language) Corpora (ISLA, Lisboa, 2008).

Augusto Soares Silva (2008a). O corpus CONDIV e o estudo da convergência e divergência entre variedades do português. Em: Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, páginas 25-28. Linguateca.

Augusto Soares Silva (2008b). Integrando a variação social e métodos quantitativos na investigação sobre linguagem e cognição: para uma sociolinguística cognitiva do português europeu e brasileiro. *Revista de Estudos da Linguagem*, 16(1):49–81.