Projecto AC/DC: acesso a corpora / disponibilização de corpora

Diana Santos & Luís Sarmento

Linguateca

Este artigo tem como objectivo apresentar o projecto AC/DC, uma iniciativa da Linguateca[1] para melhorar significativamente o acesso a corpora em língua portuguesa.

Começamos por indicar um pouco da sua história e por apresentar a motivação subjacente ao seu arranque, fornecendo uma pequena apresentação da utilidade dos corpora em linguística. Apresentamos então o seu estado presente, e as perspectivas de futuro. Finalizamos com um estudo dos nossos utilizadores, que eventualmente poderá identificar avenidas de progresso.

História

Um dos maiores entraves ao florescimento do processamento computacional do português, cedo identificado em Santos (1999a) por ocasião dos debates para o Livro Branco (1999), era a inexistência, para o português, de recursos linguísticos disponíveis para o estudo e desenvolvimento de ferramentas que processassem a nossa língua, ao contrário do que era o caso, por exemplo, para o inglês.

O projecto AC/DC constituiu-se como uma primeira resposta a esse problema, tornando todos os corpora já existentes cujos proprietários autorizavam a sua divulgação acessíveis através de uma simples interface na rede (Web). Veja-se Santos (1999b) para uma primeira proposta nesse sentido.

Contudo, era evidente que o tamanho e a qualidade dos recursos podiam ser aumentados e fomos, em paralelo com a actividade de disponibilizar os já existentes, criando outros com material de outras fontes e outros géneros, mais do que quadruplicando em tamanho as fontes iniciais.

Numa primeira fase, apenas harmonizávamos, tanto quanto possível, a atomização (ou seja, a definição das unidades básicas – palavras, frases, parágrafos, às vezes títulos...); mas muito em breve, em parceria com Eckhard Bick do projecto VISL, passámos a atribuir, também, uma anotação morfossintáctica à maioria dos corpora, processo esse cujo início se encontra documentado em Santos e Bick (2000).

A segunda fase, de compreensão e validação dos corpora anotados que tínhamos,

---

[1] A Linguateca, www.linguateca.pt, é a sucessão natural do projecto Processamento Computacional do Português, que deu origem a um centro de recursos distribuído para a língua portuguesa. As actividades da Linguateca não se desenrolam apenas no SINTEF em Oslo, mas também no Departamento de Informática da Universidade do Minho em Braga, no Laboratório de Engenharia da Linguagem (LabEL) em Lisboa, e no Centro de Linguística da Universidade do Porto (CLUP) no Porto.

assim como de extensa documentação sobre esses mesmos recursos, não invalidou contudo o nosso desejo de não fornecer apenas uma interpretação sintáctica, mas sim estender o convite, permanentemente em aberto, a todos os investigadores com ferramentas automáticas ou semiautomáticas de análise morfossintáctica, de anotar os nossos corpora e darmos-lhes acesso através do projecto AC/DC.

De qualquer maneira, as nossas ambições foram sempre aumentando, de modo que surgiram dois projectos que pretendiam colmatar as duas principais limitações do AC/DC, que passamos a descrever:

A primeira limitação era o fornecermos apenas o direito de interrogar e procurar nos corpora, mas não de os processar independentemente ou transformar, o que cerceava claramente os engenheiros da linguagem (um dos nossos públicos alvo). Donde criámos dois corpora de grandes dimensões que, além de serem disponibilizados através do projecto AC/DC, também podiam ser distribuídos fisicamente a todos os que os desejassem: o CETEMPúblico e o CETENFolha, com texto jornalístico do jornal o PÚBLICO nos anos 1991-98, e do jornal a Folha de São Paulo no ano de 1994. Veja-se, sobre o primeiro, Rocha e Santos (2000) e Santos e Rocha (2001).

A segunda limitação era o facto de a anotação sintáctica, por ser totalmente automática, ter uma margem de erro apreciável, que era preciso tentar medir -- veja-se a esse respeito Santos e Gasperin (2002) – e corrigir. Daí termos iniciado, novamente em parceria com Eckhard Bick e o projecto VISL, o projecto Floresta Sintá(c)tica – veja-se Afonso et al. (2002a; 2002b).

Ambos os projectos mencionados contribuíram valiosamente para melhorar, também, o próprio projecto AC/DC, não só refinando as ferramentas usadas, como permitindo descobrir problemas e melhorar a sua documentação.

Na terceira fase, em que nos encontramos, o projecto AC/DC concentrou-se principalmente em melhorar a usabilidade do sistema, através da detecção de outros tipos de consultas às quais faria sentido dar uma resposta mais directa. Assim, já não é “apenas” uma interface comum, mas um conjunto de serviços que fornecem várias interfaces e funcionalidades para interrogar de várias maneiras os corpora subjacentes. Um serviço de corpora “à la carte” (Rocha, em prep.) está de momento em embrião.

Uma das principais limitações, a que parcialmente tentamos dar resposta com o presente artigo, é a falta de informação directamente utilizável por pessoas com pouca ou nenhuma experiência de corpora e que podem sentir-se intimidadas ou perdidas no manancial de informação e documentação que oferecemos.

Porquê o uso de corpora?

Dado que o presente artigo foi apresentado numa sessão subordinada ao tema “Recenseamento e disponibilização de informação linguístic@”, convém talvez salientar que o adjectivo *linguística* é vago entre a acepção referente à disciplina Linguística, e a acepção referente à língua, tal como em *capacidades linguísticas* ou

diversidade linguística, e que o projecto AC/DC é uma fonte poderosa e facilmente alcançável de informação linguística nesta segunda acepção. A Linguateca também tem como uma das suas principais actividades a manutenção de um catálogo e um portal actualizado sobre o processamento computacional do português, além da organização de avaliações conjuntas na área, mas o presente artigo não se debruçará sobre estas (veja-se, respectivamente, Oksefjell e Santos (1998) e Santos (2000) para a primeira vertente e Santos e Rocha (este volume) para a segunda).

O projecto AC/DC permite a um linguista interrogar interactivamente um corpus (ou vários corpora) sobre várias questões não triviais (para mais pormenores sobre as possibilidades oferecidas veja-se a secção de "Exemplos" abaixo). Um corpus é um conjunto de textos especialmente escolhidos e geralmente marcados com informação linguística ou metalinguística (exemplos típicos de ambas são, respectivamente, uma análise do próprio texto, e a atribuição a este de classificações como género literário, autor, variante, nível de língua, etc.). Na época da informática, um corpus é, além disso, servido, manipulado, varrido através de um sistema de codificação de corpora, que permite obter resultados de forma rápida e adaptada aos desejos mais frequentes dos utilizadores (veja-se Santos e Ranchhod (1999) para uma descrição contrastiva de dois sistemas de manipulação de corpora). No projecto AC/DC, usamos o sistema Corpus Workbench desenvolvido pelo IMS da Universidade de Estugarda, o IMS-CWB, descrito em Christ et al. (1994) e Evert (2002). Finalmente, um terceiro componente é a interface aos corpora propriamente ditos, ou seja, a adaptação e personalização do sistema a um contexto de uso particular (no nosso caso, na rede, com um dado conjunto de corpora portugueses, adicionando um conjunto de funcionalidades). Este último trabalho foi feito pela Linguateca, e em geral deverá ser feito por aqueles que desejam fornecer um dado serviço.[2]

Mas ainda não explicámos ao incauto linguista que nos lê, porque poderá beneficiar do acesso a material textual escrito por outros falantes que não ele... além disso, marcado com informação que pode permitir a separação do trigo e do joio. Simplesmente, podemos começar por dizer que é um auxílio à memória – não nos lembramos imediatamente de todas as acepções de uma dada palavra, de todas os contextos sintácticos regidos por um dado verbo, de todas as expressões idiomáticas contendo a palavra *mão*... Um corpus suficientemente grande permite refrescar-nos a memória. Em muitos casos, irá mesmo mostrar-nos coisas que não sabíamos. Mas não só nem sempre sabemos tudo, como muitas vezes não concordamos com os outros falantes sobre a maneira de nos exprimirmos. Um linguista descritivo terá de aceitar ouvir e ler os seus "compatriotas de língua" usando formas que ele não usa ou não gosta; mesmo um linguista normativo terá de (se quiser ter algum impacto) se informar sobre elas. Um engenheiro da linguagem terá vantagem em saber o que tem de analisar

[2] Para uma comparação entre vários tipos de distribuição de corpora na rede, veja-se Santos (1998), onde esta tripartição foi inicialmente sugerida.

para obter um dado resultado, enquanto que um simples curioso pode divertir-se olhando para a língua de modos que uma obra impressa não permite.

Donde, um corpus pode funcionar como uma obra de referência, mas é muito maior o âmbito das suas potencialidades: de facto, pode ser um auxiliar precioso na investigação sobre a língua, se um linguista pensar em hipóteses e as testar, e, ao ver o resultado, as refinar, e tornar a interrogar o corpus como um oráculo... A questão da associação é, a esse respeito, muito pertinente: uma pergunta leva a outra, um teste leva a uma nova hipótese e a outros testes, lentamente os factores associados a um dado fenómeno podem vir a aparecer (porque é que a maior parte dos meus resultados vêm em frases negadas?) ou a serem considerados insignificantes (afinal o facto de o sujeito ser um pronome pessoal masculino em 50% dos casos não é significativo, visto que isso é o caso em geral....)

Muitas pessoas poderão ter a ideia de que tudo isto está muito bem para os outros, mas exactamente para o seu problema de investigação não precisam – não há qualquer informação que possa ser obtida ou discutida com um corpus. Na maior parte dos casos, após alguma reflexão, é possível verificar que algo pode ser obtido em conversa com um corpus... Tem a certeza? Ou falta um corpus do tipo de linguagem que está a estudar? Que tal contactar-nos?

Descrição do material textual

O projecto AC/DC é um projecto dinâmico, com alguns corpora crescendo diariamente e com a intenção de ir adicionando outros corpora que venham a ser compilados por outros investigadores ou por nós. Dado isso, uma tabela com a sua composição está fatalmente datada. As tabelas apresentadas referem-se, assim a Janeiro de 2003. A tabela 1 descreve a origem textual dos variados corpora e a 2 o seu tamanho, enquanto a tabela 3 apresenta informação sobre a distribuição das classes gramaticais em vários corpora anotados de português de Portugal, de acordo com o analisador sintáctico descrito em Bick (2000). As fontes, os detentores dos direitos de autor, e o processamento individual a que cada corpus foi submetido, encontram-se disponíveis no sítio do AC/DC.

Tabela 1. Descrição sucinta de cada corpus

natura / natpanot	Texto jornalístico, PÚBLICO, Portugal, 1991-1994, dois parágrafos por edição
enpcpub / enpcanot	Literatura traduzida do inglês, de 5 obras do ENPC
minho / minhanot	Texto jornalístico regional: Diário do Minho, Portugal, antes da revisão
eci-ebr / ebranot	Texto brasileiro, do corpus Borba-Ramsey, compilado pelo ECI
eci-ee / eeanot	Texto de chamada do programa europeu ESPRIT, em português de Portugal, compilado pelo ECI
saocarlos / scanot	Texto do corpus NILC, em português brasileiro, contendo

	maioritariamente texto jornalístico, mas também cartas comerciais e textos didácticos
frasespp / fppanot	Frases em português de Portugal
frasespb / fpbanot	Frases em português do Brasil
cetempublico(prmi) /cpprmianot/ cetempanot	Texto jornalístico, PÚBLICO, Portugal, 1991-98, extractos de dois parágrafos, baralhados.
ancib / ancibanot	Correio electrónico em português brasileiro correspondente ao tráfego na lista ANCIB, de 1998 a 2002
diaclav / diaclavanot	Texto jornalístico regional: Diário de Coimbra, Diário de Leiria, Diário de Aveiro, Viseu Diário, Portugal, 1999-2000
avante / avantanot	Texto jornalístico partidário, Avante!, Portugal, 1997-2002
classlppe	Textos literários de escritores da língua portuguesa (poesia, prosa e teatro), séculos XVI a XIX

Tabela 2. Dados quantitativos sobre os corpora AC/DC

Nome do corpus	Tamanho em unidades	Tamanho em palavras	Tamanho em frases
NATURA	7.257.152	6.257.950	225.673
NATPANOT	7.321.642	6.268.817	225.734
ENPCPUB	89.864	72.244	4.369
ENPCANOT	90.574	72.392	4.369
MINHO	2.083.761	1.738.475	53.040
MINHANOT	2.107.826	1.747.274	53.185
ECI-EBR	891.655	722.012	45.530
EBRANOT	898.542	723.007	44.689
ECI-EE	30.157	26.515	780
EEANOT	31.127	27.140	780
SAOCARLOS	41.372.756	32.091.996	1.955.166
SCANOT	41.917.324	32.378.253	1.951.484
FRASESPP	19.340	16.225	594
FPPANOT	19.542	16.208	594
FRASESPB	22.486	19.155	651
FPBANOT	22.730	19.165	651
CETEMPUBLICOPRMI	1.198.015	997.695	38.151
CPPRMIANOT	1.202.938	995.851	38.251
ANCIB	642.671	515.364	20.640
ANCIBANOT	576.450	460.442	18.306
DIACLAV	7.441.109	6.488.273	228.856
DIACLAVANOT	7.529.495	6.549.823	210.741
AVANTE	7.607.651	6.488.201	204.686

AVANTANOT	7.685.242	6.512.510	204.833
CLASSLPPE	1.872.381	1.307.334	74.174
CETEMPUBLICO	229.038.019	191.687.833	7.082.094
CETEMPANOT	229.861.093	195.730.593	7.081.564
Total[3] não anotado	**299.567.017**	**248.429.272**	**9.934.404**
Total anotado	**299.264.525**	**251.501.475**	**9.935.181**

Tabela 3. Classes gramaticais (PoS) em alguns corpora

Categoria gramatical	CETEM Público	CPPRMI ANOT	ENPCAN OT	MINHAN OT	AVANT ANOT	DIACLA VANOT
% substantivos	22,30	23,98	19,29	24,51	24,48	23,01
% verbos	14,34	14,17	19,03	14,06	13,33	15.30
% adjectivos	6,27	6,73	5,27	6,34	7,29	5,73
% pronomes pessoais	1,67	1,52	4,57	1,34	1,57	1,50
% preposições	19,16	20,45	15,77	20,88	20,33	19,87
% conjunções	4,39	4,13	5,49	4,67	5,19	4,70
% advérbios	5,76	5,41	8,04	4,74	5,32	5,79
% determinantes	19,91	20,75	17,81	20,87	21,15	20,39
% especificadores	1,90	1,80	2,48	1,68	1,98	1,99
% numerais	2,36	2,61	1,13	2,82	1,74	2,20

Descrição da informação associada aos corpora

Além da documentação associada a cada corpus, ou seja, a descrição em português da sua origem, do processamento que lhe foi aplicado, dos eventuais problemas que lhe estejam associados, da versão corrente, etc., todos os corpora têm três tipos de informação associada, a qual pode ser usada e interrogada através do serviço:

São eles atomização, marcação de características extralinguísticas, e anotação morfossintáctica. Nenhuma destas categorias é garantidamente correcta, ou seja, é possível que essa informação não tenha sido correctamente atribuída, mas trabalhamos continuadamente no sentido de as melhorar e de informar sobre os problemas ainda não resolvidos.

A nossa convicção, contudo, é a de que o fornecimento dessa informação pode ajudar bastante o utilizador, mesmo que nela não confie cegamente. Há, além disso, sempre a possibilidade de não se entrar em conta com ela. Ficamos particularmente gratos, contudo, se problemas ainda não descritos – e, como tal, provavelmente ainda não detectados – nos forem comunicados.

[3] O material do PÚBLICO do corpus NATURA encontra-se no CETEMPúblico (visto que este engloba todo o material publicado pelo Público), assim como o corpus CETEMPUBLICOPRMI é o primeiro milhão do corpus CETEMPúblico, com uma revisão manual da separação de frases. Há, contudo, razões para manter os corpora independentes, não só históricas mas também para permitir análises das diferenças.

Tipos de perguntas

Como mencionado acima, o utilizador tem ao seu dispor um conjunto de interfaces ou serviços que lhe permitem extrair diferente tipo de informação dos corpora, todos eles relacionados com os conceitos de concordância, distribuição, distribuição cruzada, frequência e número de ordem (‘ranking’).

Assim, pode procurar simultaneamente em todos os corpora (ou seleccionar um grupo) e comparar a frequência relativa e/ou a distribuição cruzada; ou pode interrogar um corpus de cada vez, ou pode ainda obter um resultado ordenado por frequência, assim como o número de ordem.

Em qualquer caso, a seguinte informação encontra-se disponível para cada unidade de um corpus: forma, lema, categoria gramatical, tempo verbal ou caso pronominal, pessoa ou número, género, função sintáctica, e informação sobre derivação automática. Além disso, é possível ter algumas palavras combinadas numa expressão com várias palavras, assim como identificar (e seleccionar) apenas as que fazem parte de um título (por exemplo).

Exemplos de utilização

Os leitores são vivamente instados a experimentar directamente com o sistema e a navegar ajudados pelos vários exemplos presentes no sítio do AC/DC – fontes suplementares são Santos e Ranchhod (1999) e Santos (2002). Em qualquer caso, apresentamos aqui alguns resultados para “abrir o apetite”.

*Procura: ".*ação". Pedido: Distribuição das **formas**. Corpus: CETEMPúblico 1.7 2008862 casos. Distribuição: Houve 7318 formas diferentes. Apresentam-se apenas ...*

situação	93174
relação	60704
criação	39226
informação	36087
população	35764
organização	35348
Associação	35015
administração	33060
formação	32393
participação	29624
operação	26014
realização	25199 etc.

Procura: "caras". Distribuição de **pos**. Corpus: CETEMPúblico 1.7 anotado 2.0 10569 casos. Distribuição: Houve 5 tipos de pos diferentes.*

N	7793
ADJ	2553
ADJ_n	203

PROP	19
ADV	1

*Procura: [lema=”ver”]. Distribuição de **temcagr**. Corpus: Natura/Diário do Minho anotado v. 2.4. 844 casos. Distribuição: Houve 14 tempos ou casos diferentes.*

infinitivo	381
particípio passado	160
presente do indicativo	92
perfeito do indicativo	89
presente do conjuntivo	40
imperfeito do indicativo	25
perfeito ou mais que perfeito	21
gerúndio	11
futuro do indicativo	8
mais que perfeito simples	6
imperfeito do conjuntivo	5
condicional	4
imperativo	1
futuro do conjuntivo	1

Procura: camarada. Distribuição de frequência do lema camarada. Frequência relativa por 10.000 palavras

AVANTANOT	1,688
EBRANOT	0,304
ENPCANOT	0,138
CETEMPANOT	0,111
NATPANOT	0,091
CPPRMIANOT	0,080
DIACLAVANOT	0,056
MINHANOT	0,046
SCANOT	0,038

*Procura: [lema="branco"]. Distribuicao de **func.** Corpus: CETEMPúblico 1.7 anotado 2.0. 22709 casos.*

adjectivo pós-modificador de nome (N<)	12138
núcleo de SP (P<)	6104
núcleo de SN sujeito pré-verbal (SUBJ>)	951
núcleo de predicativo do sujeito (<SC)	770
pós-modificador de nome, mais longe (N<PRED)	595
núcleo de SN objecto (<ACC)	579
núcleo de SN em frase sem verbo (NPHR)	404
adjectivo pré-modificador de nome (>N)	257
núcleo de SN sujeito pós-verbal (<SUBJ)	122 etc.

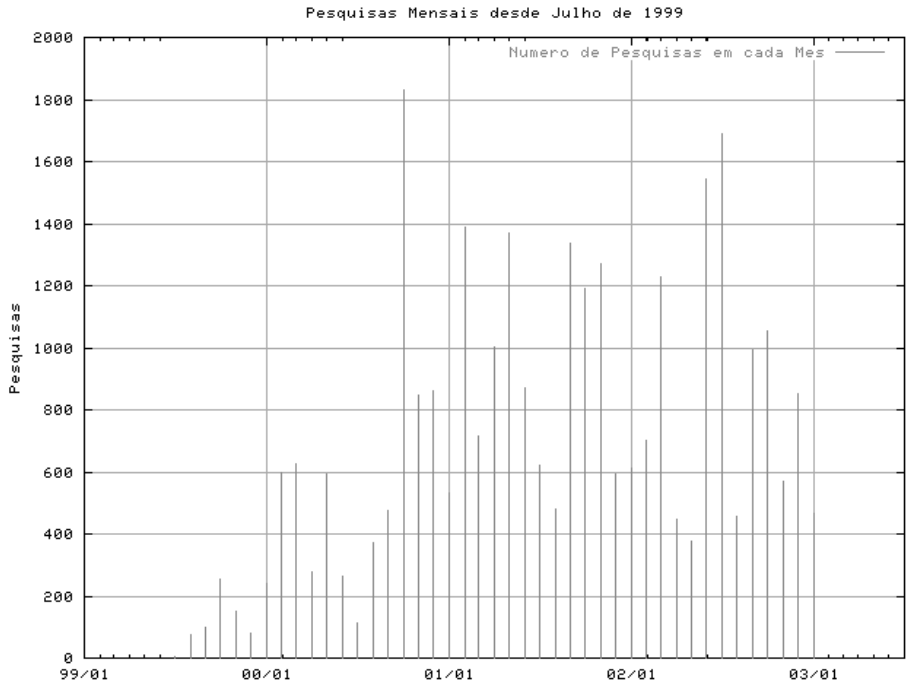
Procura: [lema="vender"] [func!="<ACC" & pos!="V."]* @[func="<ACC" & pos="N"] Distribuição de **lema**. Corpus: CETEMPúblico 1.7 anotado 2.0. 11793 casos. Distribuição: Houve 1607 lemas diferentes. Apresentam-se apenas ...*

milhão	595
produto	499
acção	355
posição	235
exemplar	234
participação	228
terreno	194
bilhete	183
parte	180
livro	173
arma	168 etc.

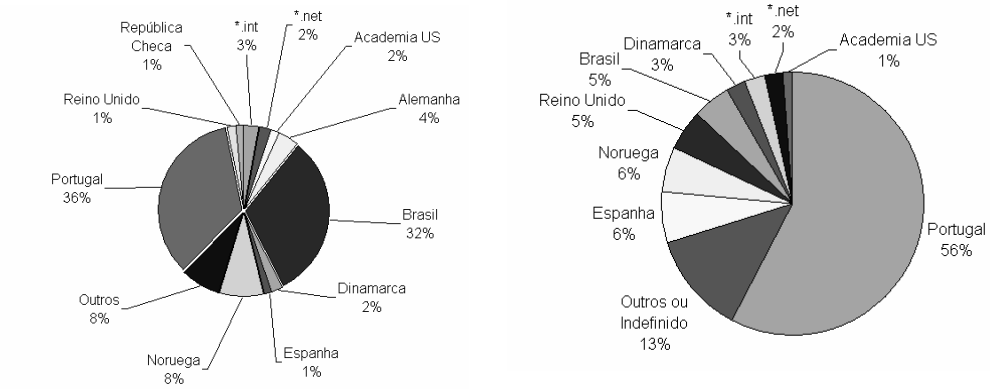
Perfil do utilizador do AC/DC

Ao contrário da maior parte dos recursos / produtos que um utilizador recebe ou compra, o projecto AC/DC assume-se como um serviço. Isto quer dizer que damos resposta e tentamos ajudar todos os nossos utilizadores. Quer também dizer que os monitoramos para ter uma ideia do que está a falhar e dos problemas com que se podem deparar. Ultimamente a preocupação com a usabilidade na informática tem crescido e com ela a actividade de estudos não intrusivos dos utilizadores na rede; cf. Burton e Walther (1999), He et al. (2002).

Pensamos ser os primeiros a publicar estudos dessa natureza sobre serviços de acesso a corpora na rede. Brines-Moya e Hartill (1998) debruçam-se sobre outro tipo de avaliação destes serviços.



Usando os programas descritos em Sarmento (2003), calculámos a distribuição mensal de acessos ao AC/DC (sem CETEMPúblico), que apresentamos na figura anexa, assim como identificámos vários valores interessantes: 23% dos utilizadores fizeram apenas uma pesquisa e nunca mais voltaram; enquanto 45% dos nossos utilizadores fizeram dez ou mais pesquisas. 41% das pesquisas feitas até agora não produziram resultados, embora a média de resultados devolvidos seja de 1252. Os utilizadores fiéis (que tenham efectuado mais de 15 consultas ao sistema), correspondendo a 2811 sessões, têm em média um tempo de 11 minutos por sessão e executam 8,77 pesquisas numa sessão. Das 30.106 pesquisas sobre as quais incidiu a nossa análise, 9882 (33%) incluiram expressões regulares.



Quanto à distribuição geográfica, os acessos de Portugal ao AC/DC sem CETEMPúblico (figura da esquerda) são apenas ligeiramente mais numerosos do que os do Brasil, enquanto que se registou um número de acessos considerável de países europeus (Alemanha e Dinamarca à cabeça) e de instituições educativas americanas. No caso do CETEMPúblico (figura da direita), e por se tratar de português de Portugal, há dez vezes mais acessos do que do Brasil.

Agradecimentos

Estamos gratos a todos os membros da equipa do projecto AC/DC que participaram em maior ou menor grau na construção e melhoria do projecto: por ordem alfabética, Alexsandro Santos Soares, Luís Costa e Paulo Rocha, assim como a todos os possuidores de corpora que gentilmente nos deixaram disponibilizá-los, e a todos os utilizadores que se aproximaram com perguntas, sugestões e dúvidas, ou que simplesmente utilizaram o serviço que concebemos para eles.

Referências

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. 2002a ""Floresta sintá(c)tica": a treebank for Portuguese", in Manuel González Rodríguez & Carmen

Paz Suárez Araujo (eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29-31 May 2002), ELRA, pp. 1698-1703.

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. 2002b "Floresta sintá(c)tica: um treebank para o português", in Anabela Gonçalves & Clara Nunes Correia (orgs.), *Actas do XVII Encontro da Associação Portuguesa de Linguística* (Lisboa, 2-4 de Outubro de 2001), APL, pp. 533-45.

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

Burton, Mary C. & Joseph B. Walther. "The value of Web log data in use-based design and testing", in *Make IT Easy 2001 "Innovate for Simplicity"* (San Jose, California, 4-7 June 2001).

Brines-Moya, Natalia & Julie Hartill. 1998 "Criteria for user-oriented Evaluation of Monolingual Text Corpora", in Antonio Rubio, Natividad Gallardo, Rosa Castro and Antonio Tejada (eds.), *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998), Vol. 2, pp.893-98.

Christ, Oliver. 1994 "A modular and flexible architecture for an integrated corpus query system", in *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research* (Budapest, July 7-10, 1994), pp.23-32.

Evert, Stefan. 2001 "CQP Query Language Tutorial", IMS Stuttgart, 13 Out 2001.

He, Daqing, Ayse Göker & David J. Harper. 2002 "Combining evidence for automatic Web session identification", *Information Processing and Management* **38** (2002), pp.727-47.

Livro Branco do Desenvolvimento Científico e Tecnológico Português (1999-2006) 1999 Observatório das Ciências e das Tecnologias, Ministério da Ciência e da Tecnologia, Portugal.

Oksefjell, Signe & Diana Santos. 1998 "Breve panorâmica dos recursos de português mencionados na Web", in Vera Lúcia Strube de Lima (ed.), *Anais do Terceiro Encontro de Processamento da Língua Portuguesa (Escrita e falada), PROPOR'98* (Porto Alegre, 3-4 novembro 1998), pp.38-47.

Rocha, Paulo Alexandre & Diana Santos. 2000 "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa", in Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, (Atibaia, São Paulo, Brasil, 19 a 22 de Novembro de 2000), pp.131-140.

Rocha, Paulo. em preparação "Tecelão: corpora a la carte".

Santos, Diana. 1998 "Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts", in Antonio Rubio, Natividad Gallardo, Rosa Castro and Antonio Tejada (eds.), *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998), pp.475-481.

Santos, Diana. 1999a “Processamento computacional da língua portuguesa: documento de trabalho”. SINTEF, Oslo, versão base de 9 de Fevereiro; revista a 13 de Abril, http://www.linguateca.pt/branco/.

Santos, Diana. 1999b "Disponibilização de corpora através da WWW", in Palmira Marrafa & Maria Antónia Mota (orgs.), *Linguística Computacional: Investigação Fundamental e Aplicações: Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística* (Lisboa, 25-27 de Maio de 1998), Colibri, pp. 323-346.

Santos, Diana & Elisabete Ranchhod. 1999 “Ambientes de processamento de corpora em português: Comparação entre dois sistemas”, in *Actas do IV Encontro sobre o Processamento Computacional da Língua Portuguesa (Escrita e Falada), PROPOR* (Évora, 20-21 de Setembro 1999), pp. 257-268.

Santos, Diana & Eckhard Bick. 2000 "Providing Internet access to Portuguese corpora: the AC/DC project", in Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), ELRA, pp.205-210.

Santos, Diana. 2000 "O projecto Processamento Computacional do Português: Balanço e perspectivas", *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR’2000)*, Atibaia, São Paulo, Brasil (19 a 22 de Novembro de 2000), pp.105-113.

Santos, Diana & Paulo Rocha. 2001 "Evaluating CETEMPúblico, a free resource for Portuguese", *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (Toulouse, 9-11 July 2001), pp.442-449.

Santos, Diana & Caroline Gasperin. 2002 "Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation", in Manuel González Rodríguez & Carmen Paz Suárez Araujo (eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29-31 May 2002), ELRA, pp.597-604.

Santos, Diana. 2002 "*Med* e *com*: um estudo contrastivo português – norueguês", *XV Skandinaviske romanistkongress* (Oslo, 12.-17. august 2002), *Romansk Forum* **16** (2) (2002), pp. 1029-42, http://www.digbib.uio.no/roman/Art/Rf-16-02-2/por/Santos.pdf.

Santos, Diana & Paulo Rocha. este volume “AvalON: uma iniciativa de avaliação conjunta para o português”.

Sarmento, Luís. 2003 “Biblioteca de análise de logs: descrição e funcionalidades”, Linguateca, CLUP, Porto, http://poloclup.linguateca.pt/biblioteca_analise_logs/ .