CQPWeb: uma nova plataforma de pesquisa para o CRPC

Amália Mendes, Michel Généreux, Iris Hendrickx, Luísa Pereira,

Maria Fernanda Bacelar do Nascimento, Sandra Antunes

Centro de Linguística da Universidade de Lisboa

Resumo

We present a newly available online resource for Portuguese, a new version of the Reference Corpus of Contemporary Portuguese, now searchable via a user-friendly web interface. We report on work carried out on the corpus previous to its publication online, namely how the corpus was built, our choice of metadata and the processes and tools involved for the cleaning, preparation and annotation to make the corpus suitable for linguistic inquiries. We also describe the web platform and resume the extensive search options available for linguistic or NLP studies.

Palavras-chave

Corpus, Cleaning, Linguistic Preprocessing, Linguistic Inquiries

Corpus, Limpeza, Pré-processamento Linguístico, Pesquisa *Online*

1. Introdução

O Corpus de Referência do Português Contemporâneo (CRPC)[1], desenvolvido no Centro de Linguística da Universidade de Lisboa (CLUL) [2], está agora disponível numa nova plataforma de pesquisa *online*, constituindo um recurso de exploração de *corpora* essencial à comunidade científica para estudos nas áreas da linguística e do processamento da língua natural (PLN).

Considerando que a análise linguística de *corpora* de grandes dimensões está dependente da flexibilidade e facilidade de acesso da tecnologia em que se baseia, foi dada especial atenção, não só à escolha da nova plataforma de acesso, mas também à preparação do *corpus*, limpando-o de toda a informação considerada irrelevante, enriquecendo-o com anotação morfossintática e lematização e convertendo-o para um formato adequado à sua publicação *online*.

Seguidamente, apresentaremos uma descrição da estrutura do *corpus* (secção 2), das tarefas necessárias para a sua disponibilização, como é o caso da organização dos

[1] http://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc
[2] http://www.clul.ul.pt/

metadados (secção 2.1), das ferramentas utilizadas para limpeza (secção 2.2) e da anotação e lematização automáticas (secção 2.3), da escolha da plataforma de acesso *online* (secção 3) e das diferentes opções de pesquisa oferecidas, com exemplificação de casos paradigmáticos (secção 4).

2. Constituição do *corpus* CRPC

O Corpus de Referência[3] do Português Contemporâneo contém, atualmente, cerca de 312 milhões de palavras (310M escrito; 1,6M oral).

A constituição deste *corpus*, tendo como objetivo fornecer informação abrangente sobre o português contemporâneo, inclui amostragens de discurso escrito (literário, jornalístico, técnico, científico, didático, económico, jurídico, parlamentar e *varia*) e oral (elocuções informais e formais). Estas amostragens dizem respeito a variedades nacionais e regionais do português, incluindo essencialmente o português europeu, mas também as variedades brasileira, de África (Angola, Cabo Verde, Guiné-Bissau, Moçambique e São Tomé e Príncipe) e da Ásia (Macau e Timor-Leste). Do ponto de vista cronológico, os textos vão desde a segunda metade do século XIX até 2008, sendo, na sua maior parte, posteriores a 1970 (Bacelar do Nascimento *et al.*, 2000).

Para atingir a diversidade referida, os procedimentos foram evoluindo ao longo do tempo. Os primeiros trabalhos, baseados principalmente na digitalização (*com OCR*) de diversos tipos de documentos, exigiram muito tempo e recursos humanos e permitiram que o *corpus* escrito tenha alcançado uma substancial cobertura em termos de tipos textuais. A atual dimensão do *corpus* mostra que foi sendo sucessivamente alargado, beneficiando, mais recentemente, dos avanços tecnológicos que têm vindo a ocorrer. Também o *corpus* oral pôde ser incrementado no âmbito de vários projetos, sendo o mais recente o C-ORAL-ROM[4]. Este projeto europeu consistiu na criação de um *corpus* comparável de quatro línguas românicas (espanhol, português, francês e italiano), com cerca de 300.000 palavras para cada língua, cobrindo discurso formal e informal. Nesta fase, o *corpus* oral ainda não está disponível para pesquisa na plataforma agora apresentada.

---

[3] O termo "corpus de referência" é usado para indicar que o corpus foi planeado para fornecer informação abrangente sobre o português contemporâneo, não sendo apresentado como uma referência de uso.

[4] http://www.clul.ul.pt/en/research-teams/189-c-oral-rom-integrated-reference-corpora-for-spoken-romance-languages

Apresentam-se, de seguida, os *corpora* que foram desenvolvidos ao longo dos trabalhos de compilação do CRPC ou que contribuíram para a sua constituição interna, e a forma como estão disponibilizados:

- Sem restrições:
  - Português Fundamental[5]
  - Português Falado[6]
- Pesquisa *online*:
  - *Corpus* de Referência do Português Contemporâneo[7]
  - *Corpus* CINTIL[8] (colaboração com NLX[9])
  - *Corpus* África[10]
- No catálogo ELRA:
  - *Corpus* PAROLE[11] (colaboração com INESC-ID[12])
  - *Corpus* C-ORAL-ROM[13]
  - *Corpus* CINTIL (colaboração com NLX)[14]

O CRPC, que já se encontrava disponível para consulta *online* desde 2002, numa versão com um *subcorpus* de 11,4M de ocorrências, foi presentemente remodelado, estando agora o registo escrito disponível na sua totalidade (310M). É de salientar que, ao constituir o *corpus*, se optou por inserir o máximo de documentos possível, embora tal facto possa contribuir para um maior desequilíbrio a nível de registos, nesta fase.

Nos Quadros 1 e 2, em baixo, resumem-se algumas das suas atuais características:

Tipos de Texto	% Docs.	% Palav.	Nº de Palavras
Jornal	50,8	35,70	110.503.376
Texto parlamentar	45,9	52,70	163.267.089
Revista	1,4	2,40	7.581.850
Varia	1,2	1,60	4.806.176
Texto jurídico	0,3	0,94	2.927.953
Livro	0,3	6,60	20.557.296
Correspondência	0,03	0,03	88.370
Folheto	0,01	0,03	80833
Total	99,94	100,00	309812943

Quadro 1: Distribuição por tipos de texto

[5] http://www.clul.ul.pt/en/resources/84-spoken-corpus-qportugues-fundamental-pfq-r
[6] http://www.clul.ul.pt/en/resources/83-spoken-portuguese-geographical-and-social-varieties-r
[7] http://alfclul.clul.ul.pt/CQPweb/
[8] http://cintil.ul.pt/
[9] http://nlx.di.fc.ul.pt/
[10] http://www.clul.ul.pt/en/resources/82-online-queries-to-crpc-subcorpora-corpus-query-tool-concor-r2
[11] http://catalog.elra.info/product_info.php?products_id=765
[12] http://www.inesc-id.pt/
[13] http://catalog.elra.info/product_info.php?products_id=757
[14] http://catalog.elra.info/product_info.php?products_id=1102

Variedade	% Docs.	% Palav.	Nº Palavras
Portugal	93,3	93,50	289.840.619
Angola	5,5	3,50	10.744.627
Cabo Verde	0,3	0,46	1.449.269
Macau	0,3	0,70	2.086.763
Moçambique	0,2	0,40	1.126.299
São Tomé e Príncipe	0,2	0,20	537.600
Brasil	0,2	1,10	3.539.770
Guiné-Bissau	0,04	0,10	364.421
Timor	0,0008	0,04	123.575
Total	100,0408	100,00	309.812.943

Quadro 2: Distribuição por variedades

Cada documento do *corpus* está associado a uma base de dados hierarquizada que contém toda a informação descritiva considerada relevante para a sua identificação, tendo por base o tipo de texto (livro, jornal, revista, etc.). Deste modo, enquanto um texto de jornal incluirá informação sobre o título do jornal, a secção a que pertence, a data, etc., num texto literário os campos dos metadados incidirão sobre o título da obra, o nome do autor, a editora, a data de edição, o país de edição, o país de origem do autor, entre outros. Uma vez que o *corpus* inclui textos de diferentes variedades do português, cobrindo um período de tempo muito vasto, considerou-se que seria interessante fornecer a possibilidade de restringir as pesquisas por estes campos de metadados, e poder selecionar, por exemplo, autores nascidos em Portugal, com português europeu como primeira língua, mas que vivem em Moçambique, estando, por isso, as suas obras registadas no *corpus* como pertencendo a essa variedade do português. Para tal, tornou-se necessário adaptar a esta nova plataforma de acesso os campos de metadados considerados relevantes.

3. Outros *corpora* de português

Iremos nesta secção dar informação sobre outros *corpora* de grandes dimensões existentes para o português. O *CETEMPúblico*, com cerca de 190 milhões de palavras, é um *corpus* constituído por notícias do jornal *Público* (Rocha & Santos, 2000) e está anotado ao nível morfossintáctico e sintáctico. O *corpus* está acessível através do

projecto AC/DC no site da Linguateca[15] através do mesmo programa subjacente à plataforma CQPWeb, o IMS Open Corpus Workbench (CWB). Outro *corpus* de grandes dimensões é o *Corpus do Português*, que contém cerca de 45 milhões de palavras e foi constituído com base noutros *corpora* de português europeu e de português do Brasil. A interface de acesso[16] permite pesquisas muito completas. Para o português do Brasil, foi constituído o *Banco do Português* (Berber Sardinha, 2003), com cerca de 230 milhões de palavras, dos quais cerca de 1 milhão pode ser pesquisado *online*[17]. Existe ainda o *Corpus NILC*, com 32 milhões de palavras, acessível através do site da Linguateca, no âmbito do projecto AC/DC. O projecto Lácio-Web[18], uma continuação do *Corpus NILC*, teve como objectivo a constituição de diversos *corpora* acessíveis *online*, num total de cerca de 10 milhões de palavras (Aluísio *et al*, 2004).

4. Preparação e anotação do *corpus* CRPC

Para permitir a facilidade de acesso ao *corpus* e a compatibilização com a nova plataforma, foi necessário desenvolver vários tipos de tarefas, que seguidamente se descreverão.

4.1. Limpeza do *corpus*

A limpeza do *corpus* foi outra das tarefas que mereceram particular atenção. O facto de o CRPC conter documentos obtidos a partir de fontes diversas tornou necessário proceder-se à sua limpeza automática, sobretudo no que diz respeito às secções de publicidade, *spam* ou outra informação irrelevante presentes nos documentos. Assim, começou por ser criado um *corpus* anotado manualmente para construção de um *corpus* de treino, fornecendo materiais que permitem à ferramenta de limpeza NCLEANER (Evert, 2008) aprender a distinguir entre segmentos relevantes e segmentos não relevantes. Este *corpus*, de 200 documentos (cerca de 200,000 palavras) escolhidos aleatoriamente a partir do total de documentos do CRPC, foi tratado em seguida com a referida ferramenta, para se eliminarem secções não relevantes. Com este

[15] http://www.linguateca.pt/ACDC/
[16] http://www.corpusdoportugues.org/
[17] http://www2.lael.pucsp.br/corpora/bp/conc/
[18] http://www.nilc.icmc.usp.br/lacioweb/

trabalho, foi atingida uma redução de 433 para 309,8 milhões de palavras (cerca de 28%).

2.2. Anotação morfossintática e lematização

Para a tokenização do *corpus* foi utilizada a ferramenta LX-tokenizer (Branco & Silva, 2004), que, ao remover pontuação e detetar fronteiras de frases, dá conta, entre outros, de fenómenos como a contração de formas ou o reconhecimento de clíticos nas várias posições.

Seguiu-se a anotação morfossintática, com o etiquetador MBT (Daelemans *et al.*, 1996), que usou, como base, o *corpus* de treino CINTIL[19], de 1M de palavras (Barreto *et al*., 2006). Para os casos não identificados, foi feita uma anotação manual com categorias morfossintáticas e lemas. A anotação, de momento seguida para o *corpus* escrito, baseou-se em 80 etiquetas morfossintáticas, com categorias principais (Quadro 3). As principais diferenças entre a anotação do CRPC e a do CINTIL dizem respeito às unidades multilexicais (que, nesta fase, foram removidas na sua grande maioria), às formas contraídas (que são mantidas como tal e anotadas com duas etiquetas (por exemplo, das\PREP+DA)) e às etiquetas de flexão nominal e verbal (género, número ou pessoa), que ainda não estão contempladas no CRPC. A anotação automática atinge 95,5% de correcção.

Etiqueta	**Categoria**	**Exemplos**
ADJ	Adjetivos	*bom, brilhante, eficaz*
ADV	Advérbios	*hoje, já, sim, felizmente*
CN	Nomes Comuns	*computador, cidade, ideia*
DA	Artigos Definidos	*o, os*
IA	Artigos Indefinidos	*uns, umas*
DEM	Demonstrativos	*este, esses, aquele*
CL	Clíticos	*o, lhe, se*
CARD	Cardinais	*zero, dez, cem, mil*
DGT	Dígitos	*0, 1, 42, 12345, 67890*
ORD	Ordinais	*primeiro, centésimo, penúltimo*
IND	Indefinidos	*tudo, alguém, ninguém*

[19] http://cintil.ul.pt/cintilfeatures.html

INT	Interrogativos	*quem, como, quando*
EXC	Exclamativos	*que, quanto*
POSS	Possessivos	*meu, teu, seu*
PRS	Pessoais	*eu, tu, ele*
CJ	Conjunções	*e, ou, tal como*
PREP	Preposições	*de, para, em redor de*
INF	Infinitivo	*ser, afirmar, viver*
INFAUX	Infinitivo de verbo auxiliar	*ter, havermos*
VAUX	Forma finita de "ter" ou "haver" em tempos compostos	*temos, haveriam*
V	Verbos (não PPA, PPT, INF e GER)	*falou, falaria*
PPT	Particípio Passado em tempos compostos	*tinha afirmado, tinha vivido,*
PPA	Particípio Passado em tempos não compostos	*livros lidos*

Quadro 3: Etiquetas de anotação do *corpus* escrito

Por último, procedeu-se à lematização do corpus. Foi adaptada a ferramenta MBLEM (Van den Bosch & Daelemans, 1999) que combina um léxico de lemas e formas do português (desenvolvido no CLUL no âmbito do projeto Dicionário Eletrónico do Português) com um algoritmo de aprendizagem. Com algumas ferramentas, o processo de lematização consiste numa verificação no dicionário (por exemplo TreeTagger (Schmidt, 1994) ou o etiquetador de Bick (Bick, 2000) e as palavras desconhecidas não são lematizadas. Pelo contrário, a ferramenta MBLEM combina uma verificação sobre um dicionário e um algoritmo de aprendizagem treinado sobre o dicionário para também dar conta da lematização de palavras não reconhecidas. O lematizador obteve um valor de 96.7% de correcção sobre uma amostra do CINTIL com 50.000 palavras, contendo 17,117 formas plenas (as palavras gramaticais não são lematizadas).

Assim, no âmbito deste trabalho foram adaptadas ao português duas ferramentas para o processamento de corpora, nomeadamente um etiquetador morfossintático e um lematizador (para mais pormenores, consultar Généreux *et al*, 2012).

3. Plataforma *online*

Por ser o maior e o mais diversificado *corpus* do português a ficar disponível para consultas *online*, a escolha da melhor plataforma de acesso ao CRPC mereceu particular

atenção. Deste modo, foram consideradas quatro ferramentas de interface: CQPWeb[20], Glossa[21], Manatee[22] e Glozz[23]. Um estudo comparativo destas quatro candidatas permitiu-nos concluir que o CQPWeb (Hardie, em prep.) seria a que melhor se adaptava às necessidades de processamento de consultas (tendo em conta a linguagem de pesquisa utilizada – *Corpus Query Processor*) e a que possuía uma melhor facilidade de navegação para os utilizadores do recurso. Utilizando o sistema operativo UNIX, a plataforma requer um servidor Apache com suporte MySQL, PHP, Perl e Open Corpus Workbench[24].

O CRPC está disponível para pesquisa *online* (http://alfclul.clul.ul.pt/CQPweb), como acima referido. O utilizador pode optar entre um acesso sem registo[25] ou com registo[26]. A versão registada (que apenas requer o preenchimento de um formulário *online* para obtenção do nome de utilizador e da palavra-chave) tem como principais funcionalidades a possibilidade de o utilizador criar *subcorpora* com base nos metadados, compilar e descarregar listas de frequências para cada *subcorpus* e guardar no servidor do CLUL as consultas efetuadas e os *subcorpora* criados para posteriores consultas. Esta plataforma fornece, igualmente, uma extensa lista de possibilidades de pesquisa (descritas no manual do utilizador[27]). Seguidamente, apresentaremos algumas dessas opções, realçando a importância do uso deste recurso em diversos estudos e projetos.

4. Potencialidades do recurso

A consulta ao CRPC pode ser feita através de vários tipos de pesquisa, que pode ser simples (procurando em todo o *corpus* uma palavra particular, como 'livro', 'casa', etc.) ou restrita, em que é possível selecionar a variedade (Portugal, Brasil, etc.) e o tipo de discurso (jornal, livro, revista, etc.), bem como especificar padrões de pesquisa para a extração de concordâncias, nomeadamente através de expressões regulares, sequências de palavras, lemas, classes morfossintáticas e elementos contraídos. A Figura 1, em

[20] http://cqpweb.lancs.ac.uk
[21] http://tekstlab.uio.no/glossa//html/GLOSSA_manual.html
[22] http://www.textforge.cz/products
[23] http://www.glozz.org
[24] http://cwb.sourceforge.net/
[25] http://alfclul.clul.ul.pt/CQPweb/
[26] http://alfclul.clul.ul.pt/CQPnet/
[27] http://alfclul.clul.ul.pt/CQPweb/doc/CRPCmanual.v1_en.pdf

baixo, ilustra a pesquisa do lema ‘poder’, enquanto nome comum (que inclui as formas ‘poder’ e ‘poderes’) em textos de português de Portugal, provenientes de jornais.

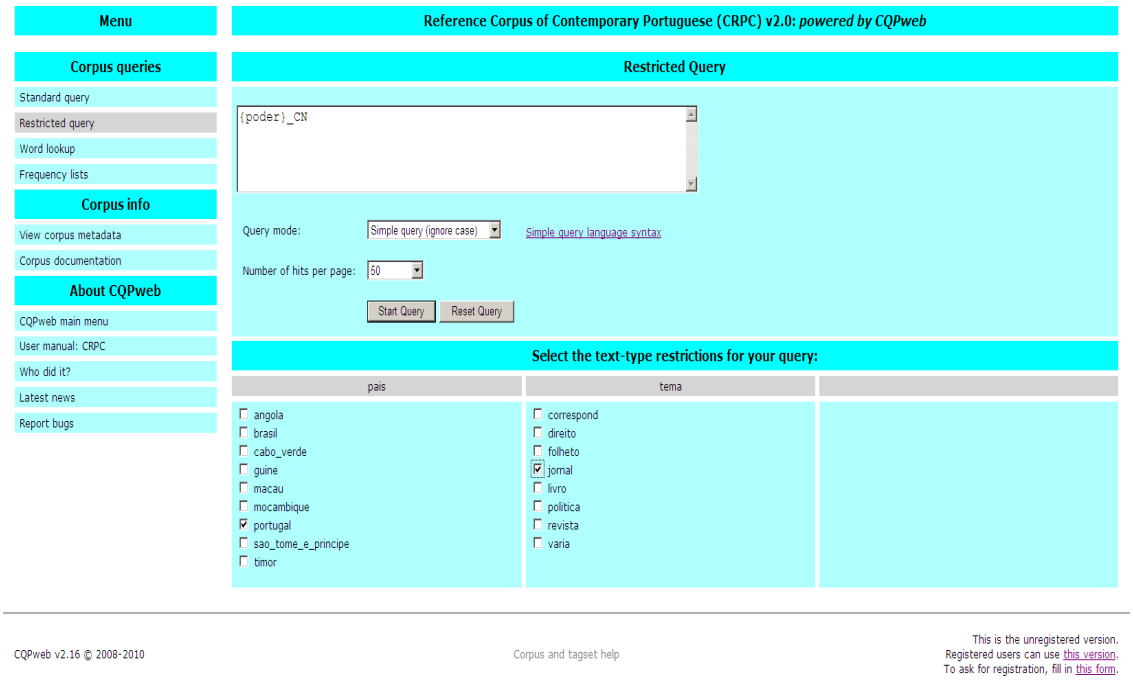

Figura 1: Pesquisa do lema nominal ‘poder’ com opção *restricted query*

O Quadro 4 ilustra alguns tipos de pesquisas combinadas que podem ser efectuadas.

Tipo de pesquisa	**Exemplo**	**Exemplos de correspondência**
Expressões regulares	+mente	*absolutamente, provavelmente*
	lind[o,a]*	*lindo, lindos, linda, lindas, lindamente*
Classes morfossintáticas	_IND	*algo, nada, ninguém, outras*
	ante*_V	*antecipar, antever, antedatar, antepor*
Lemas	{poder}	*poder, posso, podes, podia, pudesse*
Sequências de palavras	*_ADJ {jantar}_CN	*célebre jantar, breve jantar, grandes jantares*
	{de} +**jantar	*de estar presente num jantar, de fazer um jantar*
Elementos contraídos	{em\+*}	*no, nos, na, nas, naquele, naquela, neste*

Quadro 4: Pesquisas combinadas no CRPC

Após a obtenção do resultado da pesquisa, estão ainda disponíveis várias opções (cf. Figura 2). É possível, por exemplo, ordenar as concordâncias. Por defeito, estas aparecerão ordenadas alfabeticamente pela primeira palavra à direita da palavra-nó, mas é possível alterar essa ordenação até 5 palavras à esquerda ou à direita da palavra pesquisada.

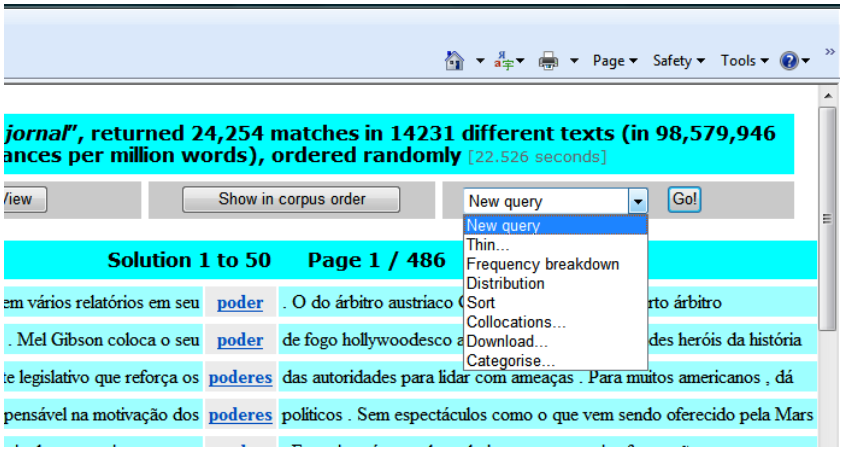

Figura 2: Opções da plataforma CQPweb

No caso dos utilizadores registados, a opção *Keywords* permite comparar listas de frequências que podem ter sido compiladas com base em diferentes *subcorpora*, sendo possível identificar automaticamente formas que apenas ocorrem num determinado *subcorpus*, o que, por sua vez, facilita a realização de estudos contrastivos entre variedades do português ou diferentes tipos de texto. Na área do PLN, esta função pode constituir um recurso importante no treino e desenvolvimento de ferramentas para o processamento do português e, mais particularmente, das variedades nacionais e dos diferentes tipos de discurso.

Ao consultar o *corpus*, e depois de obter uma concordância, é igualmente possível obter informação adicional sobre associações de palavras através da função *Collocations*, que apresenta várias opções de pesquisa. A Figura 3 ilustra a informação obtida (quer a nível de co-ocorrentes quer a nível estatístico) através da consulta da palavra 'janela', tendo-se selecionado a medida estatística de associação lexical log-likelihood, uma distância de até 3 palavras à direita e à esquerda entre a palavra pesquisada os seus co-ocorrentes e com o mínimo de 5 ocorrências no *corpus*.

Collocation controls

Collocation based on:	Word form	Statistic:	Log-likelihood
Collocation window *from*:	3 to the Left	Collocation window *to*:	3 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: (none)	Submit changed parameters Go!

There are 7,462 different words in your collocation database for "[word="janela"%c]". (Your query "janela" returned 5,283 matches in 2421 different texts) [34.006 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood value
1	pela	459,379	47.001	606	460	1991.248
2	uma	1,971,066	201.666	1,077	799	1882.461
3	ABERTA	22,282	2.28	238	206	1744.769
4	da	4,537,364	464.233	1,480	764	1433.505
5	à	10,957,196	1121.067	2,476	1075	1274.909
6	quarto	18,750	1.918	126	107	807.36
7	Abrir	15,686	1.605	99	89	622.026
8	Indiscreta	96	0.01	39	26	586.634
9	abriu	8,028	0.821	70	53	484.527
10	parapeito	158	0.016	35	31	475.353
11	vidro	3,971	0.406	58	52	460.675
12	para	2,629,825	269.066	666	504	418.273

Figura 3: Associações obtidas para a palavra 'janela'

Além dos estudos lexicais que estas consultas permitem, a possibilidade de avaliar resultados de acordo com diferentes medidas lexicais (informação mútua, t-score, z-score, log-likelihood, etc.) poderá também constituir um recurso importante na área do PLN.

O CRPC foi já utilizado em muitos estudos e projetos[28], salientando-se, entre os mais recentes, um estudo lexical tendo como base *subcorpora* comparáveis das variedades africanas do português (Bacelar do Nascimento *et al.*, 2008), uma análise lexical de base estatística do Diário da Assembleia da República, no período que antecedeu e seguiu a revolução de 1974 (Généreux *et al.*, 2010), o estudo das propriedades dos verbos leves em predicados complexos (Duarte *et al.*, 2009) e uma proposta de anotação de valores modais (Hendrickx *et al.*, 2012).

[28] http://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc

5. Conclusão e trabalho futuro

Este artigo permite dar a conhecer o trabalho envolvido na preparação e disponibilização *online* do CRPC, focando os processos de limpeza e anotação do *corpus*, bem como a escolha da plataforma de acesso e o modo como pode ser facilmente usada (por fornecer um ambiente amigável para o utilizador e possuir um alargado leque de possibilidades de pesquisa), contribuindo para o desenvolvimento de estudos linguísticos e a criação de recursos na área do tratamento automático de línguas. Pretende-se, em futuras versões: (i) proceder a uma segunda fase de limpeza do *corpus*, que se centrará em aspectos de segmentação; (ii) aumentar e melhorar a constituição interna do CRPC, tornando-o mais equilibrado; (iii) alargar os campos de metadados pesquisáveis; (iv) acrescentar, na anotação, etiquetas de flexão nominal e verbal, bem como informação sintática e classificação das combinatórias; (v) introduzir um detetor de línguas para identificar algumas palavras estrangeiras presentes no *corpus*.

A equipa do CRPC pretende ainda assegurar, na medida do possível, autorizações de editores e autores, com a finalidade de disponibilizar textos integrais de parte do *corpus*.

Desde a sua publicação *online*, em Março de 2011, a plataforma de acesso ao CRPC tem sido visitada e consultada por utilizadores de todo o mundo, numa média de 1600 pesquisas por mês.

Referências

Aluísio, S., G.M. Pinheiro, A.M.P. Manfrim, L.H.M. de Oliveira, L. C. Genoves Jr., S.E.O. Tagnin (2004) The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisboa, Portugal, pp. 1779-1782.Bacelar do Nascimento, Maria Fernanda, Luísa Pereira & João Saramago (2000) Portuguese Corpora at CLUL. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*. Atenas, Grécia: vol. II, pp. 1603-1607.

Bacelar do Nascimento, Maria Fernanda, Antónia Estrela, Amália Mendes & Luísa Pereira (2008) On the use of comparable corpora of African Varieties of Portuguese

for linguistic description and teaching/learning applications. In *2nd Workshop on Building and Using Comparable Corpora (LREC)*. Marraquexe, Marrocos: pp. 39-46.

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Bacelar do Nascimento, Filipe Nunes & João Ricardo Silva (2006) Open Resources and Tools for the Shallow Processing of Portuguese: The TagShare Project. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. Genoa, Italy: pp. 1438-1443.

Berber Sardinha, Tony (2003) The Bank of Portuguese. DIRECT Papers 50. LAEL, PUCSP.

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

Branco, António & João Silva (2004) Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisboa, Portugal: pp. 507-510.

Daelemans, Walter, Jakub Zavrel, Peter Berck & Steve Gillis (1996) MBT: A memory-based part of speech tagger-generator. In *Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora*, pp. 14-27.

Duarte, Inês, Madalena Colaço, Anabela Gonçalves, Amália Mendes & Matilde Miguel (2009) Lexical and syntactic properties of complex predicates of the type light-verb + noun. *Arena Romanistica 4*, pp. 48-57.

Evert, Stefan (2008) A Lightweight and efficient tool for cleaning webpages. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*. Marraquexe, Marrocos.

Généreux, Michel, Amália Mendes, Luísa Pereira & Maria Fernanda Bacelar do Nascimento (2010) Lexical analysis of pre and post revolution discourse in Portugal. In *3rd Workshop on Building and Using Comparable Corpora (LREC)*. La Valletta, Malta: pp. 65-71.

Généreux, Michel, Iris Hendrickx, Amália Mendes (2012) A Large Portuguese Corpus On-Line: Cleaning and Preprocessing. In Caseli, H. et al. (eds.) *Computational*

Processing of the Portuguese Language. Proceedings of the 10th International Conference PROPOR1012. Berlin, Heidelberg: Springer-Verlag, pp. 113-120.

Hardie, Andrew (em prep.) CQPweb - combining power, flexibility and usability in a corpus analysis tool.

Hendrickx, Iris, Amália Mendes & Mencarelli, Silvia (2012) *Modality in Text: a Proposal for Corpus Annotation*. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*. Istambul, Turquia: pp. 1805-1812.

Rocha, Paulo & Diana Santos (2000) CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In Nunes, Maria das Graças Volpe (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.

Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Van den Bosch, Antal & Walter Daelemans (1999) Memory-based morphological analysis. In *Proceedings of the 37th Annual of ACL*, pp. 285-292.