CORPUS DE REFERÊNCIA DO PORTUGUÊS CONTEMPORÂNEO (CRPC)
- Desenvolvimento e Aplicações -

Maria Fernanda Bacelar do Nascimento
José Bettencourt Gonçalves
Centro de Linguística da Universidade de Lisboa

O **Corpus de Referência do Português Contemporâneo (CRPC)** é um projecto em curso no CLUL, que tem como responsável o Prof. Doutor João Malaca Casteleiro e como coordenadora Maria Fernanda Bacelar do Nascimento. A equipa é actualmente constituída ainda por Maria Lúcia Garcia Marques, Investigadora do CLUL, por 6 colaboradores em linguística e 3 informáticos, todos a tempo parcial. Instituições Financiadoras do CRPC: Fundação Calouste Gulbenkian (Serviço de Educação); Junta Nacional de Investigação Científica e Tecnológica (JNICT) – Programa Estímulo em Ciências Sociais e Humanas; Caixa Geral de Depósitos (mecenato); e uma rede de Fornecedores de dados[1].

O **Corpus de Referência do Português Contemporâneo (CRPC)** é um *corpus* que se pretende que venha a ser representativo do português de uso geral e corrente do séc. XX, sendo, actualmente, constituído por amostragens de língua falada e de língua escrita do português europeu, do português do Brasil, do português de países africanos de língua oficial portuguesa e do português de Macau.

Está assim em curso o estabelecimento de uma base de dados e de conhecimentos constituída por documentos linguísticos organizados e informatizados, já acessíveis aos investigadores, professores, tradutores e a todos os que em Portugal e no estrangeiro desejem aceder

a dados linguísticos atestados para realizarem trabalhos de carácter teórico ou prático em que intervenha a língua portuguesa.

O CRPC contém, presentemente, 30 milhões de palavras informatizadas; mais de 1 milhão de palavras de discurso oral (espontâneo e formal) e 29 milhões do discurso escrito (literário, jornalístico, didáctico, técnico, científico, publicitário, etc.); além do português europeu, estão hoje representadas no *corpus*, como já foi dito, outras variantes nacionais e constituiram-se *sub-corpora* especializados: jurídico, político, económico, religioso, informático, da astronomia, da moda, da decoração, e outros. Com o desenvolvimento do Projecto, prevê-se que até 1998 o *corpus principal* venha a atingir os 50 milhões de palavras.

O Projecto CRPC foi aprovado pelo antigo Instituto Nacional de Investigação Científica (INIC) e teve início em 1988. Dispondo, inicialmente, de recursos muito escassos, a equipa começou por se dedicar essencialmente à recolha bibliográfica para estabelecimento das fontes (literárias, jornalísticas, didácticas, técnicas e científicas), tarefa com que se iniciou a constituição do *corpus* escrito. O Projecto foi apresentado a várias Instituições e, graças ao bom acolhimento que o Serviço de Educação da Fundação Calouste Gulbenkian lhe concedeu, atribuindo-lhe, em 1992 um subsídio considerável, mais tarde a um financiamento da Junta Nacional de Investigação Científica e Tecnológica (JNICT) no âmbito do Progama Estímulo em Ciências Sociais e Humanas, e por fim, a outras instituições que o subsidiam ou fornecem dados, foi possível incrementá-lo nos últimos três anos.

A criação do CRPC justificava-se pela necessidade de elaborar novas e actualizadas descrições do português, quer no seu uso "médio", quer nas suas variedades, para o que se tornava absolutamente indispensável a existência de um *corpus de referência* cuja centralização, organização e informatização, permitisse uma rápida e fácil consulta dos dados seleccionados.

A construção de um *corpus* desta natureza apresenta diversos problemas de ordem prática e teórica:

Os problemas de ordem prática ligam-se, essencialmente, à difícil tarefa de obviar aos enormes custos que acarretam a manutenção de uma equipa de colaboradores regulares, a aquisição de equipamento, a informatização (por digitação, OCR ou conversão), correcção, revisão, codificação e desenvolvimento de software adequado ao tratamento e análise do *corpus*.

Os problemas de carácter teórico que o CRPC suscita são essencialmente os que respeitam ao *Desenho do Corpus, Representação dos Textos, Extracção de Informação* e *Anotação.*

Quando da concepção do Projecto, traçou-se um primeiro esquema do *corpus*. Contudo, a continuação do estudo da Bibliografia sobre o tema, publicada nos últimos 30 anos, o conhecimento de outros *corpora* por participação directa nos Projectos europeus em que veio a colaborar a equipa do CRPC: Network of European Reference Corpora (NERC), Preparatory Action for Linguistic Resources Organization for Language Engineering (PAROLE), Expert Advisory Group on Language Engineering Standards (EAGLES), e o próprio desenvolvimento do CRPC têm influenciado a tomada de decisões sobre a dimensão, o desenho e o equilíbrio do *corpus*.

Presentemente o CRPC é um *corpus* aberto, ou seja, um *corpus* de dimensão sempre crescente e composição muito heterogénea, apto a fornecer indicações sobre o "estado da língua". Deste *corpus principal* podem extrair-se, de acordo com as necessidades dos utilizadores, *sub-corpora* mais reduzidos, adequados em dimensão e desenho à realização de análises exaustivas, objectivamente perspectivadas.

Paralelamente ao *corpus principal* constituímos ainda *sub-corpora* especializados que, contrariamente àquele, se caracterizam precisamente pela homogeneidade e servem, portanto, para ilustrar certos tipos de linguagem e não a linguagem comum.

No que respeita à língua falada, deu-se continuidade ao *corpus* oral do Português Fundamental com a recolha de amostragens de discurso informal e também formal: rádio, televisão, discurso parlamentar, aulas, conferências (sempre que os textos não são lidos). Das 700.000 palavras informatizadas do Português Fundamental, passámos actualmente para cerca de 1 milhão e meio de palavras que incluem transcrições de português europeu, do Brasil, da Guiné, de Cabo Verde e de Macau; o Arquivo Sonoro do CRPC contém mais de 600 horas de gravação e foi recentemente valorizado com gravações audiovisuais de discurso parlamentar, cedidas pela Assembleia da República.

O *corpus* escrito, como dissemos, contém cerca de 29 milhões de palavras e é constituído por amostragens de textos literários, jornalísticos, didácticos, técnicos, científicos, publicitários, correspondência, e outros.

No caso dos textos literários, por razões que se prendem com a utilização do *corpus* em lexicografia, foi feito um inventário de autores

e obras de língua portuguesa a partir da segunda metade do séc. XIX, submetido depois à apreciação de especialistas e que, mais recentemente, está a ser completado com o apoio do Serviço de Bibliotecas e Apoio à Leitura da Fundação Calouste Gulbenkian.

Também a selecção de textos pertencentes a outros géneros tem pretendido seguir critérios rigorosos de tipologia de *corpora* de referência, na escolha e dimensão dos fragmentos retirados das obras seleccionadas.

Dada a evidente relação entre a qualidade última dos trabalhos realizados e os dados observados, têm sido feitas múltiplas experiências e confrontos para que o *corpus* e principalmente os *sub-corpora* analisados resultem equilibrados (cfr. BACELAR DO NASCIMENTO, M. F., 1994); importa, contudo, dizer que a actual composição do CRPC não resulta somente da aplicação das metodologias de selecção e recolha de dados, mas resulta em grande medida também das possibilidades de acesso a materiais e dos meios humanos e financeiros disponíveis.

Quanto à *Representação dos textos* foram definidos vários níveis de representação dos textos orais e escritos, níveis que dizem respeito, quer à codificação para utilização humana, quer à codificação para transmissão electrónica de informações.

O *corpus* é constituído por uma série de textos ou por grupos de textos com uma proveniência comum (por exemplo, artigos de um mesmo jornal). Para todos são fornecidas, em bases de dados, informações que os identificam, classificam (título, género, fonte, autor, data, etc.) e dão conta do estado electrónico do documento (digitação ou leitura óptica; corrigido; revisto). São estabelecidas ligações entre as bases de dados e o *corpus*.

Quanto ao próprio texto, diversas soluções têm sido tomadas. No que respeita ao *corpus* oral, foram ensaiados dois tipos de representação gráfica do sinal sonoro, divergentes, principalmente, quanto à utilização ou não de sinais de pontuação gráfica, marcação de divergências auditivas e de sobreposição de vozes, problemas teoricamente sempre em aberto (Cfr. BACELAR DO NASCIMENTO, M. F. *et alii*, 1987; BACELAR DO NASCIMENTO, M. F. *et alii*, 1989; CASTELEIRO, J. M. e M. F. BACELAR DO NASCIMENTO, 1994). Quanto ao *corpus* escrito, foi produzido um documento de trabalho, "Normas para Tratamento do Texto" (cfr. PEREIRA, L. A. S., 1993) em que são

especificados todos os procedimentos e codificações relativas a uma multiplicidade de informações sobre, por exemplo, paginação, notas, citações, símbolos numéricos ou fórmulas, abreviaturas, etc.

Os *corpora* que conhecemos têm adoptado esquemas de representação muito diversos, o que prejudica a comparação de resultados. Em virtude da integração do CRPC em redes europeias, estamos a iniciar a adaptação das nossas normas aos *standards* internacionais.

As principais ferramentas informáticas desenvolvidas no CLUL por João Miguel Casteleiro permitem a extracção automática de informações como: Frequência, Repartição e Concordâncias permitindo estas uma grande flexibilidade na sua utilização.

Quanto à *Anotação Morfossintáctica*, a convite do Professor Geoffrey Leech, da Universidade de Lancaster, foram feitas a verificação da adaptabilidade ao português da *Anotação Morfossintáctica* proposta pelo EAGLES e subsequentes sugestões já integradas na última versão das recomendações para anotação morfossintáctica daquele projecto comunitário.

Posteriormente, o Grupo de Linguagem Natural do INESC, nosso parceiro no Projecto PAROLE, fez a adaptação do seu analisador morfológico PALAVROSO à codificação do EAGLES. Está actualmente em curso a validação deste esquema de anotação morfossintáctica sobre amostragens do CRPC.

Quanto à *Anotação Sintáctica*, também a convite do Departamento de Linguística da Universidade de Lancaster, comentou-se e verificou-se a relevância para o português do sistema de anotação proposto pelo Subgrupo de Anotação Sintáctica do Grupo de Trabalho sobre *Corpus* do EAGLES.

Foi feito o estudo do processo de anotação proposto e a sua aplicação a uma pequena amostragem do *corpus* do português escrito, tendo sido enviado ao EAGLES um texto com propostas de solução para aspectos específicos do português.

O CRPC tem desempenhado um papel importante em actividades de docência universitária e na realização de dissertações, em que é tomado como objecto de estudo (colaboração particularmente importante com a Faculdade de Letras da Universidade de Lisboa) e, no desenvolvimento de outros projectos de investigação nacionais e internacionais, quer no âmbito do CLUL, quer no âmbito das relações de colaboração que se foram estabelecendo com muitas outras instituições.

Para além das participações em projectos europeus já citados e da colaboração com projectos brasileiros e da África lusófona, as aplicações do CRPC têm sido mais relevmantes em áreas como a lexicografia, em que é tomado como fonte de selecção de entradas lexicais, como fonte de selecção de abonações e ainda como lugar de observação de fenómenos linguísticos como, por exemplo, o das associações lexicais; é o CRPC que está na base do *sub-corpus* desenhado para o "Dicionário de Combinatórias do Português", outro projecto do CLUL que, neste Encontro, será apresentado durante a mesa-redonda sobre dicionários.

O *corpus* oral do CRPC está também a ser utilizado num projecto de ensino do Português LE, em curso no CLUL, no âmbito do Programa LINGUA, tendo como parceiros a Universidade de Toulouse-le-Mirail e a Universidade da Provença (Aix-Marselha), projecto que se intitula: "Português Falado – Variedades geográficas e sociais".

Reconhecidas que são, quer a escassez de materiais para o ensino de português que utilizem documentos sonoros autênticos, recolhidos em diferentes situações de comunicação, quer a falta de estudos descritivos sobre o português falado, importa aqui referir, como mais uma das aplicações do CRPC, neste caso do seu *sub-corpus* oral, este projecto que consiste na publicação em CD-ROM, de amostragens de português falado, contendo alinhados o som e a correspondente transcrição ortográfica e ainda na publicação de estudos morfossintácticos, lexicais e pragmáticos resultantes de análises daquele *sub-corpus*. Ainda sobre um *corpus* de 4 milhões de palavras, também extraído do CRPC, no âmbito do programa LINGUA está a ser construído material didáctico pela equipa do projecto Morfo-sintaxe do Português: ensino assistido por computador, de que é responsável João Malaca Casteleiro e coordenadora Maria Elisa Macedo.A demonstração de uma das componentes desse material decorre na sala de demonstrações deste Encontro.

Estas são algumas das aplicações do CRPC. Referimos apenas as que estão em curso no CLUL, em projectos de investigação sobre a língua e em trabalhos interdisciplinares. O que importou aqui salientar foi a existência no CLUL de uma base de dados e de conhecimentos constituída por documentos linguísticos organizados e informatizados, já acessíveis aos investigadores, professores, tradutores e a todos que em Portugal e no estrangeiro desejem aceder a dados linguísticos atestados, para realizarem trabalhos de carácter teórico ou prático em que intervenha a língua portuguesa.

Notas

[1] São, actualmente, fornecedores regulares do CRPC as seguintes Instituições: Centro de Informática do Ministério da Justiça; Assembleia da República; Caixa Geral de Depósitos (que é também financiadora do CRPC); Editorial Verbo; Academia das Ciências de Lisboa; Jornais: Expresso, O Público, Diário de Notícias, Jornal de Notícias, Jornal do Minho, A Bola; Revistas: Grande Reportagem, ProTeste, Máxima; Agência Lusa; Sociedade Bíblica Portugal; Estação de Rádio: TSF; Centro de Documentação da JNICT; Serviço de Bibliotecas e Apoio à Leitura da Fundação Calouste Gulbenkian.

Referências

AIJMER, K. e B. ALTENBERG (eds.) (1991), *English corpus linguistics*, London, New York, Longman.

ATKINS, S., J. CLEAR, N. OSTLER (1992), "Corpus design criteria", *Litterary and Linguistic Computing*, 7, 1, pp. 1-16.

BACELAR DO NASCIMENTO, M. F. (1994), "Aplicação de análises linguísticas sobre *corpora* ao ensino do Português, LE", *Actas do 3.º Congresso Internacional do Ensino como Língua Estrangeira*, México, 1994 (entregue para publicação).

BACELAR DO NASCIMENTO, M. F., GARCIA MARQUES, M. L. e SEGURA DA CRUZ, M. L. (1987), *Português Fundamental*, vol. II – *Métodos e Documentos*, tomo I – *Inquérito de Frequência*, Lisboa, INIC, CLUL.

CASTELEIRO, J. M. e M. F. BACELAR DO NASCIMENTO (1994), *Final Activity Report*, produced as part of the *Report* of EC Project *Network of European Reference Corpora*, Lisboa, CLUL.

PEREIRA, L. A. S. (1993), "Normas para Tratamento do texto", documento de trabalho do CRPC, Lisboa, CLUL.

SINCLAIR, J. M. (ed.) (1987), *Looking up. An Account of the Cobuild Project in Lexical Computing*, London and Glasgow, Collins.