

**O CORPUS DE REFERÊNCIA DO PORTUGUÊS CONTEMPORÂNEO E OS
PROJECTOS DE INVESTIGAÇÃO DO CENTRO DE LINGUÍSTICA DA
UNIVERSIDADE DE LISBOA SOBRE VARIEDADES DO PORTUGUÊS
FALADO E ESCRITO**

Maria Fernanda Bacelar do Nascimento

Centro de Linguística da Universidade de Lisboa

Uma vez que o objectivo deste Colóquio é fazer o balanço crítico e a discussão do ponto actual das investigações sobre o português nas suas variedades, poderá ser útil dar a conhecer o Corpus de Referência do Português Contemporâneo, as diferentes formas de disponibilização deste corpus para investigação e, ainda, os principais projectos portugueses e internacionais que, com base nele, se têm vindo a realizar.

A linguística sobre corpora que já tinha uma longa tradição em certas disciplinas linguísticas, nomeadamente em lexicografia, atingiu nos anos 90 um extraordinário desenvolvimento, sendo hoje os modernos corpora electrónicos fortemente valorizados no quadro de diferentes teorias, disciplinas e abordagens linguísticas e em trabalhos interdisciplinares.

O grande desenvolvimento da informática e da capacidade dos computadores permitiu o armazenamento de enormes quantidades de dados linguísticos assim como a sua rápida exploração.

Inicialmente os modelos automáticos de análise linguística de corpora limitavam-se às indexações e à extracção de frequências de ocorrências e de concordâncias.

Entretanto, a linguística computacional foi concebendo formalismos(e métodos para a sua implementação) muito sofisticados mas cujos resultados não permitiam a construção de sistemas eficazes para o tratamento de textos reais. Por sua vez, as chamadas indústrias das línguas (em especial as editoriais) desenvolviam esforços com vista ao processamento da língua natural para aplicações práticas (dicionários, gramáticas, tradução, ensino, editoração de sumários, aplicações multilingues).

O encontro destes três interesses permitiu o desenvolvimento e a partilha de métodos comuns e de ferramentas para vários níveis de análises e descrições linguísticas feitas com base em corpora muito extensos e enriquecidos com diversos tipos de anotações.

Os modernos corpora linguísticos são seleccionados e ordenados de acordo com critérios explícitos e codificados de forma normalizada e

homogénea. O primeiro nível de anotação identifica os textos com referências bibliográficas precisas, fontes primárias utilizadas, responsabilidade editorial e formato. Os níveis de anotação de carácter propriamente linguístico dizem respeito à etiquetagem morfológica e a codificações sintácticas, semânticas, pragmáticas, etc., sendo o nível de anotação morfológico o mais geralmente praticado.

O Centro de Linguística da Universidade de Lisboa, tendo continuado a sua tradição de recolha e estudo de dados e actualizado os seus métodos, é hoje detentor de um corpus de grandes dimensões, constituído ao abrigo do Projecto "Corpus de Referência do Português Contemporâneo", (CRPC). Grande parte dos textos tem-nos sido cedida já em formato electrónico por uma rede de fornecedores de dados (Editoras, Jornais, Revistas, Instituições públicas e privadas como Estações de Rádio ou Projectos congéneres como os Projectos brasileiros NURC e PEUL), sendo outros textos informatizados no CLUL.

As motivações que estiveram na base da realização de um trabalho de tão grande fôlego na área dos recursos linguísticos centram-se na nossa forte convicção de que os corpora são uma componente essencial em projectos de investigação sobre a língua em que emerge a noção de que a variabilidade dos usos linguísticos ultrapassa a rigidez dos modelos preconcebidos e de que a evidência de fenómenos, proporcionada pela observação de grandes corpora, traz benefícios conjuntamente à linguística teórica, descritiva e aplicada.

A criação do CRPC também se justificava pela urgente necessidade de elaborar novas e actualizadas descrições do português, tanto no seu uso “médio” como nas suas variedades geográficas, sociais e discursivas, para o que se tornava indispensável um corpus de referência centralizado e de consulta acessível. A dimensão do CRPC ultrapassa já os 70 milhões de palavras e a sua constituição está sumariamente descrita adiante:

1 "Corpus de Referência do Português Contemporâneo"

Dimensão total: **77,3 milhões de palavras**
(Dados de 1997)

Escrito	Oral
75 617 671	1 725 240

Antes de 1900	1901-1970	Após 1970
1 000 000	2 600 000	73 730 911

PORTUGAL	ANGOLA	BRASIL	CABO VERDE
----------	--------	--------	------------

72 460 676	60 976	4 009 953	534 013
GUINÉ	MACAU	S. TOMÉ	MOÇAMBIQUE
46 984	7 997	12 000	210 321

JORNAL	LIVRO	REVISTA	VARIA
42 052 196	20 328 252	5 982 879	3 848 815
INTERVENÇÃO PARLAMENTAR	SUPREMO TRIBUNAL DE JUSTIÇA	FOLHETOS	CORRESPONDÊNCIA
1 482 132	1 437 667	322 538	163 192

A propósito da dimensão, que se pretende crescente, do CRPC importa justificar a necessidade destes “mega corpora”. Se tivermos em atenção que aproximadamente 70%

das palavras de qualquer texto pertencem às 700 palavras mais frequentes de uma língua e, também, que cerca de metade das palavras de um texto ocorre com Frequência 1, e se, a isto, acrescentarmos que o uso e/ou sentido mais típico de uma palavra é, em média, duas vezes mais comum do que todos os seus outros usos e sentidos (cfr. D.

WILLIS, 1990, p.VI e SINCLAIR, J. M., 1991), compreende-se a necessidade de construir corpora de milhões de palavras e, também, a necessidade de que as amostragens que os constituem não sejam pequenos fragmentos de cada texto, mas sim, excertos de grandes dimensões ou mesmo textos completos; isto não diz respeito, evidentemente,

a corpora destinados a estudos de fonética ou de fonologia (dado o número mais reduzido de itens que é preciso inventariar, nestes casos), mas sim àqueles que devem servir de base a estudos lexicais e sintáticos. E refiro-me à sintaxe porque foi precisamente o desenvolvimento de grandes corpora que levou os linguistas a aperceberem-se, com grande realismo, que não só o léxico mas também as regras da sintaxe constituem sistemas abertos, ao contrário do que queria Chomsky.

Só em grandes corpora ganham significado interpretações sobre a frequência de utilização de formas lexicais ou de certas estruturas, como, por exemplo, as que respeitam ao grau de lexicalização de fenómenos sintáticos.

São, ainda, os mega-corpora que permitem testar modelos e propor gramáticas probabilísticas (cfr. PEREIRA, F., 1993).

Contudo, tão importante como a dimensão é a constituição e o equilíbrio do corpus.

O CRPC é um corpus aberto, em constante evolução e com uma constituição muito heterogénea. Deste corpus monitor extraem-se sub-corpora gerais mais reduzidos adequados no seu desenho, dimensão e

composição, aos objectivos e níveis de análise pretendidos. Dele se podem ainda extrair sub-corpora especializados (literário, jornalístico, político, informático, económico, etc.) os quais, contrariamente ao corpus monitor, se caracterizam pela homogeneidade temática e constituem a base de estudos monográficos ou contrastivos sobre linguagens de especialidade; pretende-se vir a extrair, também, sub-corpora desenhados de acordo com critérios estritamente linguísticos, favorecendo o estudo de sub-linguagens, isto é, de funcionamentos linguísticos específicos, mas de momento, é, ainda, impossível proceder às análises necessárias à classificação dos textos sob essa perspectiva. Os materiais que constituem o CRPC estão identificados e classificados de acordo com critérios externos (informação bibliográfica, forma de acesso ao texto, estado electrónico do documento) e com critérios internos(género, tópico, etc.)

A hierarquia dos atributos de identificação e classificação dos textos foi especificamente organizada para textos orais e para textos escritos, tendo-se definido, também, vários níveis de representação que dizem respeito quer à sua codificação para utilização humana, quer à sua codificação para transmissão electrónica de informações. No respeitante ao oral, deu-se o maior relevo ao estabelecimento de convenções sobre os tipos de representação gráfica do sinal sonoro. Quanto aos textos escritos, incluem-se codificações relativas a uma multiplicidade de informações sobre, por exemplo, títulos, capítulos, linhas de um poema, paginação, notas, citações, símbolos numéricos ou fórmulas, abreviaturas, etc.

A participação em projectos internacionais mostrou a necessidade de adaptar as nossas Normas para Tratamento do Texto (PEREIRA, L.A.S., 1993) a formatos electrónicos de representação textual normalizados, o que está a ser feito para o corpus escrito, parte do qual se encontra já convertido em

S G M L (Standard Generalized Markup Language –

-I S O 8879) por ser uma linguagem extremamente poderosa que permite, por um lado, uma exaustiva caracterização e representação dos documentos e, por outro, a sua fácil manipulação em associação com as ferramentas necessárias a inúmeras aplicações previsivelmente pretendidas pelos utilizadores de corpora.

O CRPC é acessível para consultas, sendo os pedidos dos utilizadores analisados caso a caso.

A extracção de informação do CRPC e a sua disponibilização tem introduzido mudanças importantes nas perspectivas adoptadas em trabalhos académicos que orientamos (nomeadamente dissertações universitárias) e em projectos de investigação nacionais e internacionais em que o Centro de Linguística da Universidade de Lisboa participa.

Interrogando-nos sobre o valor de verdade das teorias que estão associadas ao uso de metodologias dedutivas e de dados da intuição, temos desenvolvido nestes trabalhos abordagens de tipo indutivo. É, pois, a própria natureza dos dados que nos conduz na procura de metodologias adequadas à sua análise

assim como na procura de modelos para a descrição

e interpretação coerente de fenómenos lexicais morfossintácticos e sintácticos evidenciados em contextos reais de diversos usos e variedades do português contemporâneo.

De entre os vários projectos já realizados ou em execução no Centro de Linguística com base neste corpus, passo a referir resumidamente três que me parecem de maior relevância: Dicionário de Combinatórias do Português, Português Falado, Variedades Geográficas e Sociais e LE-PAROLE.

Quero ainda salientar o importante papel desempenhado pelo Corpus de Referência do Português Contemporâneo na elaboração do Novo Dicionário da Língua Portuguesa em execução na Academia das Ciências de Lisboa. Este corpus é a principal fonte das abonações do dicionário e base de validação das suas entradas lexicais.

Dicionário de Combinatórias do Português

O objectivo deste Projecto subsidiado pela Junta Nacional de Investigação Científica e pelo Instituto Camões –Programa Lusitânia, e terminado em 1997, foi a constituição de um inventário de associações lexicais estabelecido a partir de um corpus de mais de 12 milhões de palavras extraído do Corpus de Referência do Português Contemporâneo.

Ampliando a noção firthiana de colocação, consideramos como combinatórias modelos associativos gradativos e mais ou menos padronizados que excluem a associação livre e atingem a frase

idiomática. A sua identificação é feita de acordo com processos de natureza quantitativa (Frequências de ocorrência e medidas de informação mútua) e qualitativa (observação e interpretação dos dados reais) o que permite aos seus utilizadores caracterizarem as palavras:

- na sua relação com outras palavras com as quais sistematicamente ocorrem de forma contínua ou descontínua (padrões co-ocorrênciais)
- na sua relação com traços gramaticais evidenciados pela co-ocorrência lexical. Por exemplo, determinado lema ou forma flexionada do lema ocorre sistematicamente com determinado tipo de verbo, tempo verbal ou construção sintáctica (padrões gramaticais).
- Nas suas relações extralinguísticas – enunciativas, situacionais, contextuais – já que associações fortes em determinado registo de língua podem ser associações fracas noutra registo de língua (padrões discursivos).

Sendo manifestamente impossível caracterizar as palavras segundo estes padrões por recurso à intuição e dado que eles se evidenciam com demasiada sistematicidade para serem tidos como acidentais, torna-se clara a enorme importância do seu estudo com base em corpora assim como a construção e aplicação de instrumentos de análise adequados.

Dos vários resultados a que os utilizadores têm acesso, distingo aqui o índice de 2.428.809 pares de diferentes formas lexicais, pares esses que ocorreram no corpus com Frequência igual ou superior a 2. As formas que constituem estes pares podem ter entre si distâncias de 1, 2, 3 ou 4 palavras para a esquerda ou para a direita da palavra nó (palavra em estudo).

Sobre estes pares, os utilizadores têm as seguintes informações:

- a Frequência de ocorrência da palavra-nó no corpus
- a Frequência de ocorrência do par
- a distância a que se encontram os elementos do par
- O índice combinatório do par (informação mútua sobre o comportamento dos elementos do par que identifica a sua significância no corpus)
- Os contextos restritos e alargados em que o par ocorreu
- A repartição dos contextos por tipos de discurso

(Apresentação de combinações da palavra COMPLEXO)

Julgamos que estes materiais constituem, mais do que um dicionário, um dicionário dos dicionários, dadas as informações que fornecem sobre a frequência de uso de lemas e de formas lexicais isoladamente e em associação e sobre os aspectos que evidenciam relativamente a factores morfológicos, sintácticos, semânticos e colocacionais que caracterizam padrões associativos do português contemporâneo.

A apresentação desenvolvida deste Projecto é tema da comunicação que vai ser apresentada por Luísa Alice Santos Pereira no dia 27.

Português Falado, Variedades Geográficas e Sociais

Este projecto enquadra-se no conjunto de trabalhos sobre corpora orais, em particular estudos lexicais e sintácticos, que vínhamos realizando. Foi executado com o apoio dos Programas europeus LINGUA-SOCRATES, programas para a promoção do conhecimento de línguas estrangeiras na comunidade europeia, no âmbito da Acção VB – Desenvolvimento e Intercâmbio de Materiais de Ensino. O Centro de Linguística da Universidade de Lisboa é a instituição coordenadora deste projecto e tem como parceiros associados a Universidade de Toulouse-le-Mirail e de Aix-en-Provence. Os trabalhos terminaram em Dezembro de 1997, tendo sido atingidos os objectivos programados, pelo que se encontram em fase de publicação os seguintes materiais:

- Quatro CD-ROM com 80 gravações de português falado informal e formal, amostragens que foram recolhidas ao longo dos últimos 25 anos, predominantemente nos anos 90, em Portugal, no Brasil, em todos os países africanos de língua oficial portuguesa e em Macau. Dos CD-ROM constam a gravação sonora e a correspondente transcrição ortográfica alinhadas: um cursor vai acompanhando, sobre um texto transcrito, a voz do falante.

-

(Demonstração de 3 excertos)

- Três volumes de estudos em que se apresentam resultados de análises morfosintáticas, sintáticas e pragmáticas realizadas sobre um corpus oral de quase 2 milhões de palavras de variantes do português.

Sendo o português uma das línguas menos ensinadas na União Europeia, não obstante ser uma das mais faladas do mundo, este projecto constitui um incremento para o ensino do Português LE, quer nos curricula académicos, quer em percursos profissionais em que é crescente a sua importância como veículo de comunicação nas relações internacionais. Mas os materiais que resultaram deste projecto ultrapassam largamente o ensino do Português-LE. A produção destes materiais e a tecnologia que lhes está associada teve, de facto, como

objectivo principal desenvolver as capacidades de compreensão (mas, também, de produção) do português falado. Contudo, o material não foi seleccionado tendo em vista perfis restritos de utilizadores pelo que se veio a revelar, conforme tem sido reconhecido, de grande interesse também no ensino da língua materna e, mais geralmente, em trabalhos de investigação sobre o português falado.

A publicação dos volumes com descrições de carácter lexical, sintáctico, pragmático e enunciativo sobre um corpus oral vem também colmatar uma lacuna nos estudos do português. Por falta de tempo destacarei apenas as análises quantitativas do léxico feitas sobre as variantes europeia e brasileira que potencializam interessantes estudos constrativos, e os estudos de sintaxe do oral feitos de acordo com a metodologia da análise em configuração preconizada por Cl. Blanche-Benveniste. Esta análise, sublinhando a inadequação da noção de frase ao discurso oral, quebra a linearidade das transcrições feitas em conformidade com as regras da escrita, permitindo visualizar, numa representação em grelha, certos fenómenos discursivos através da exploração dos eixos horizontal (desenvolvimentos sintagmáticos) e vertical (desenvolvimentos paradigmáticos) e demonstram que as fragmentações e outras marcas aparentemente anárquicas, típicas do discurso oral e geradoras da falta de entusiasmo com que é abordada a sintaxe da língua falada, não prejudicam, de facto, a estrutura sintáctica do enunciado que assim configurado se torna perfeitamente analisável.

(Exemplo de uma configuração sintáctica)

Meu Relatório, pág. 20

O Desenvolvimento deste Projecto é tema de uma comunicação que vai ser apresentada neste Colóquio por José Bettencourt Gonçalves, no dia 27.

LE PAROLE

LE-PAROLE é um projecto de Engenharia Linguística financiado pela Comissão Europeia (DG XIII), ao abrigo do Programa Telematics Application of Common Interest, que se situa na continuação de outros projectos comunitários em que também participámos e que foram realizados com vista ao desenvolvimento, normalização e aplicação de recursos linguísticos europeus, particularmente corpora e léxicos. Nele participam 14 países * europeus, estando a língua portuguesa representada pelo Centro de Linguística da Universidade de Lisboa como parceiro principal e o Instituto Nacional de Engenharia de Sistemas e Computadores (INESC) como parceiro associado.

O projecto consiste no aproveitamento de recursos linguísticos e informáticos disponíveis nos países europeus para a construção de corpora e léxicos segundo um modelo coerente e integrado de constituição de materiais e sua descrição mediante o uso de ferramentas comuns, o que permite facilitar as ligações multilíngues e dar resposta a um grande conjunto de aplicações no Processamento das Línguas Naturais.

Este projecto põe em relevo a importância do estudo das línguas feito com base em grandes quantidades de dados naturais. Para cada língua foram constituídos corpora de 20 milhões de palavras, harmonizados no que respeita ao seu desenho, composição e codificação. Cada corpus contém percentagens iguais de textos de livros, jornais, revistas e uma miscelânea, materiais estes na sua quase totalidade da década de 90. Todos os textos estão convertidos em SGML. Um sub-corpus de 250.000 palavras foi anotado morfossintacticamente, utilizando um conjunto de etiquetas comum a todas as línguas representadas. Para o português, a anotação tem sido feita automaticamente, através do analisador morfológico PALAVROSO desenvolvido no INESC, sendo a desambiguação semi-manual.

(Excerto do texto anotado e desambiguado)

O léxico é constituído por 20 mil entradas: 12.000 substantivos, 3.000 verbos, 3.000 adjectivos, 500 advérbios e 1.500 palavras gramaticais,

estrangeirismos, siglas e abreviaturas. Cada entrada é seguida de informação morfossintáctica e sintáctica codificada.

A componente sintáctica do léxico integra os seguintes tipos de informação: restrições com reflexo morfológico, subcategorização, restrições sobre a estrutura interna dos argumentos, ordem linear, função sintáctica dos argumentos, obrigatoriedade vs opcionalidade de realização dos argumentos e alternância.

A partir do próximo mês de Maio entra em funcionamento o programa SIMPLE no âmbito do qual se acrescentarão ao léxico informações semânticas relevantes.

Vão estar acessíveis, via INTERNET, corpora de 3 milhões de palavras de cada língua, incluindo a parte anotada e todos os léxicos.

Com esta comunicação quis dar a conhecer os recursos linguísticos disponíveis no Centro de Linguística da Universidade de Lisboa e explicitar os princípios teóricos e metodológicos em que assentam.

