

Dicionário de Combinatórias do Português do Centro de Linguística da
Universidade de Lisboa

Luísa Alice Pereira¹ e Maria Fernanda Bacelar do Nascimento²

Indicações gerais

Nome do Projecto: *Dicionário de Combinatórias do Português (DCP)*

Instituição em que o Projecto foi executado: Centro de Linguística da
Universidade de Lisboa (CLUL)

Responsável pelo Projecto: Professor Doutor João Malaca Casteleiro

Coordenadora do Projecto: Investigadora Maria Fernanda Bacelar do
Nascimento

Financiamento: Junta Nacional de Investigação Científica e Tecnológica e
Instituto Camões (Programa Lusitânia)

Introdução

A investigação sobre *corpora* possibilita a observação controlada de grandes massas de dados cuja importância permite determinar tendências muito fortes no uso da língua, o que, evidentemente, é fundamental na descrição lexicográfica.

As evidências empíricas que os *corpora* fornecem apontam para a relação

¹ Colaboradora de investigação do Centro de Linguística da Universidade de Lisboa.

² Investigadora principal do Centro de Linguística da Universidade de Lisboa.

entre as produções linguísticas e factores extralinguísticos (culturais, situacionais, ...). Por outro lado, a observação de tendências sistemáticas aponta para relações entre processos mentais e fenómenos de produção linguística. Na realidade, torna-se patente que as línguas naturais são mais regularmente padronizadas do ponto de vista sintagmático do que se pensava e o estudo de *corpora* permite identificar esses padrões. A identificação dos padrões associativos em que a palavra participa contribui também para determinar as suas acepções e usos reais.

Naturalmente, o estudo do *corpus* não permite determinar todos os usos possíveis, ou seja, nenhum *corpus* pode comprovar que determinado fenómeno linguístico não é possível. Pode, sim, determinar que usos são mais frequentes ou típicos, quer do uso geral (se a observação se faz num *corpus* geral), quer de usos especializados (se a observação se faz sobre *corpora* de linguagens especializadas). Tudo isto, naturalmente, se configura como essencial para a prática lexicográfica; pode mesmo dizer-se que o uso do *corpus* é visto cada vez menos como um suplemento dos dados da intuição e cada vez mais como uma parte fundamental da construção teórica.

Entre os vários usos dos *corpora* linguísticos em lexicografia, destacam-se o estudo das Frequências para selecção ou validação de nomenclatura, a extracção e análise de Concordâncias como fonte de Abonações e, cada vez mais, a identificação e análise de Padrões associativos.

Os padrões associativos são aqui entendidos como uma extensão da noção de Firth de colocação. Colocação, termo introduzido por Firth (1955: 190-215), consistia na caracterização de uma palavra de acordo com outras palavras que tipicamente ocorrem com ela. Ou seja, Firth despertou fortemente o interesse pelo estudo das co-ocorrências lexicais, mostrando que o aspecto relevante do significado de uma palavra é o conjunto de todas as outras palavras que com ela se combinam.

Ampliando este conceito, identificamos os padrões associativos segundo processos de natureza quantitativa e qualitativa (estes, até hoje, essencialmente intuitivos), na tentativa de caracterizar as palavras:

1. na sua relação com outras palavras com as quais sistematicamente ocorrem (modelos co-ocorrenciais);
2. na sua relação com traços gramaticais (morfo-sintáticos e sintáticos) que se evidenciam (determinada palavra ocorre sistematicamente com determinado tipo de verbos em determinado tempo verbal ou com determinada construção. Por exemplo, verbos de opinião, como julgar, pensar, achar.. ocorrem prioritariamente com completivas);
3. nas suas relações extra-linguísticas (situacionais, contextuais) que a repartição por tipo de discurso permite (associações fortes em determinado registo de língua podem ser fracas noutra registo de língua).

Assim, foi concebido o projecto *Dicionário de Combinatórias do Português*, cujo objectivo inicial era a constituição de um inventário das associações lexicais de uso mais frequente no português contemporâneo, estabelecido a partir de um *corpus* desenhado para o efeito, extraído do *Corpus de Referência do Português Contemporâneo* do CLUL. Este *corpus* contém diversos tipos de discurso falado e escrito do português europeu. As associações lexicais, que designámos por combinatórias, apresentam graus de coesão diversos: grupos totalmente *crystalizados*, *semi-crystalizados* ou apenas constituídos por *co-ocorrentes privilegiados*. Um relatório de associações lexicais assim concebido é, em si mesmo, de grande utilidade mas é-o também como fonte de informações a incluir noutros dicionários unilingues, bilingues e multilingues, na tradução tradicional ou assistida por computador, no levantamento de grupos neológicos vernáculos ou

importados, no estabelecimento de tipologias diversas (por exemplo, da complementação verbal ou nominal), no estudo contrastivo das variantes do português e, naturalmente, na programação que visa a produção e reconhecimento automático da língua natural. Quando da apresentação do Projecto, estava previsto que o Dicionário se desenvolvesse dentro das formas clássicas de observação e análise de co-ocorrências lexicais em *corpora* (Concordância, Frequência, Repartição) e desse origem a um dicionário tradicional, no seu género, como, por exemplo, o de Benson et al. (1986). Contudo, a conjugação de vários factores fez alterar as previsões iniciais: o estudo aprofundado da bibliografia mais recente sobre modelos de associações lexicais, de que destaco Church e Hanks (1990), Calzolari e Bindi (1990), Sinclair (1991a, 1991b e 1996), Smadja (1993) e Biber (1993); o contacto com outros projectos europeus; a colaboração neste projecto, aquando das suas deslocações a Portugal, do Professor John Sinclair da Universidade de Birmingham e da Professora Nicoletta Calzolari do Instituto de Linguística Computacional de Pisa; e as experimentações de novas metodologias de análise realizadas pela equipa de linguistas e informáticos que participaram neste projecto permitiram a perspectivação de objectivos qualitativa e quantitativamente mais ambiciosos. Consistiram estes novos objectivos em permitir ao utilizador o acesso informático directo a todas as combinatórias que ocorreram num *corpus* de 12.282.392 palavras (11.426.197 escrito; 856.195 oral) com diferentes posições relativamente à palavra em estudo (palavra-nó), assim como o acesso directo aos contextos em que ocorrem (contextos restritos ou alargados), Frequência de Ocorrência e Repartição por tipo de discurso e significância da combinatória no *corpus* (Índice Combinatório determinado estatisticamente), conjunto de informações a seleccionar e a manipular directamente pelos utilizadores de acordo com os seus objectivos particulares. Isto porque se considerou que o Projecto, tal como se ia redefinindo, se tornava uma fonte inesgotável de informações que permitiam ultrapassar grandemente as finalidades inicialmente previstas.

Nos trabalhos feitos sobre o Corpus de Referência de Português Contemporâneo - CRPC, para o *Dicionário de Combinatórias do Português - DCP*, ou para integração em dicionários gerais de língua, o estudo de Padrões associativos tem-se mostrado grandemente produtivo. Todos estes estudos, baseados em *corpora*, são essencialmente empíricos, procurando dar conta dos usos a partir de dados autênticos: utilizam técnicas interactivas de análise automática (informática) e manual dos dados e, por isso mesmo, associam técnicas de análise quantitativas e qualitativas (ou seja, incluem nos resultados da análise observações de natureza estatística e de natureza interpretativa). No trabalho sobre Padrões associativos, a identificação faz-se, como se disse, de uma forma empírica, a partir da análise de grandes corpora que são tidos como representativos ou da linguagem comum, seja oral ou escrita, ou de uma linguagem especializada (científica, técnica, ...).

O Corpus DCP

O *corpus* sobre o qual se processa a pesquisa de combinatórias, extraído do Corpus de Referência do Português Contemporâneo, é constituído por 11.426.197 palavras de discurso escrito e 856.195 palavras de discurso oral (*corpora* com dimensão de cerca de 10 milhões de palavras têm sido considerados capazes de assegurar a fiabilidade dos resultados na extracção de combinatórias [Smadja: 1993]). Este *Corpus* tem a seguinte distribuição:

Língua Escrita

Discurso jornalístico	60%
Discurso literário	24%
Discurso técnico, científico e didáctico	10%

Discurso político	4%
Discurso jurídico	2%

Língua Falada

Discurso espontâneo	80%
Discurso formal	20%

Análise com base na Frequência simples

Para conseguir obter padrões associativos, faz-se uma exploração do *corpus* com base na Frequência. No entanto, quando se trata de Frequência bruta, temos resultados como os que se seguem:

Excerto das frequências de palavras em -mente (por ordem decrescente)

<u>Início</u>		<u>Aos 24%</u>	
		sensatamente	10
nomeadamente	1633	risonhamente	10
finalmente	1491	rijamente	10
igualmente	1466	resignadamente	10
relativamente	1380	primorosamente	10
<u>Aos 10%</u>		<u>Aos 60%</u>	
tardiamente	43	vocacionalmente	1
simbolicamente	43	vocacionadamente	1
		vividamente	1
		vitaliciamente	1
		vistosamente	1
<u>Aos 16%</u>			

tremendamente	20
teimosamente	20
sumariamente	20

Frequências no *corpus* dos lemas dos sinónimos de *Notável* apresentados em dicionários

excelente	816
célebre	451
notável	424
famoso	413
extraordinário	335
singular	313
ilustre	201
distintos	169
apreciável	198
extravagante	60
insigne	20

Se pretendermos usar qualquer destas duas listas de frequências, ou outras do género, para decidir, por exemplo, sobre nomenclatura de um dicionário, sobre constituição de um léxico,..., teremos dificuldade em saber que escolhas fazer e como fundamentá-las. Uma boa ajuda nas decisões a tomar poderia ser a extracção de Contextos lematizados, o que permitiria uma melhor análise de natureza interpretativa.

Contextos do Lema *Notável* com ordenação -1,-2:

A ele se ficaram a dever comédias notáveis, como algumas da série @iPink «corda seca» e da «aresta». Outro conjunto notável, onde predominam os a da Bacalhoa, destacam-se dois conjuntos notáveis: o da Casa do Tanque, c conservando-se ainda muitos conjuntos notáveis, onde as duas técnicas refe espontânea e livre que praticou conseguiu notáveis realizações de tipo impre designavam a hoje "notável" Constância. Notável foi o adjectivo atribuído p

assem a submissão, tiveram contrapartidas notáveis. Nas mais formosas e sa
a sua terra, que dá um contributo notável para os tais 30 000, temos 4944 de

, passo fome e peço dinheiro emprestado! Notável, como organização de vida...
um porto flutuante, uma obra de engenharia notável para a época, de modo a que
o Oliveira. Porfírio, que efectuou uma época notável no União de Leiria, irá ter
verente nº 91 dos Bulls de fazer uma época notável. Destaque ainda para Scottie
ido por um verdadeiro espírito de equipa, notável capacidade física e alguns áv

Do excerto apresentado, constata-se, porém, que as simples chaves de
ordenação não permitem a associação concentrada de determinado co-
ocorrente junto do nó com o qual se associa. De facto, a partir da
concordância de um determinado nó, é possível fazer manipulações que
forneçam o co-ocorrente pretendido. Mas, como podemos saber que um co-
ocorrente é mais significante que outro, junto de um determinado nó?

Também uma lista em bruto de todos os pares deste *corpus*, que têm
Frequência ≥ 2 e que, à partida, poderia parecer indicação suficiente acerca
de possíveis nós e seus co-ocorrentes não é muito mais informativa, como se
pode ver no seguinte exemplo:

Excerto da lista de *pares* do nó *Notável*, com Frequência ≥ 2 :

2	2	1	101640786	efeito notável
2	2	1	101702261	eficácia notável
2	2	1	119382465	figura notável
2	2	1	121835697	forma notável
2	2	1	129071702	imaginação notáveis
2	2	1	15613625	acção notável
2	2	1	159724486	notáveis deste
2	2	1	159724954	notáveis eu
2	2	1	159725066	notáveis no

Passou-se à exploração do *corpus* com extracção de concordâncias de tipos diferentes:

- . contextos de diversas dimensões: 1, 2, 3, 4 linhas antes e depois da palavra em estudo (palavra-nó);
- . contextos de 1 linha, incluindo a palavra-nó a meio da linha e 5 palavras antes e 5 palavras depois do nó;
- . contextos ordenados de diferentes formas, quer alfabeticamente antes ou depois do nó, quer de acordo com as frequências dos co-ocorrentes.

Contudo, estes tipos de exploração não fornecem uma categorização relativa dos padrões associativos encontrados.

Informação de carácter estatístico

Com base nas propostas teóricas e nas técnicas descritas em Church e Hanks (1990) e em Sinclair (1991), procedeu-se à tentativa de caracterização dos grupos lexicais a partir de *medidas de informação mútua* que definem indicadores estatísticos sobre índices combinatórios, índices de tipicidade, distância média entre o nó e o co-ocorrente, variância e fixidez das combinatórias (Pereira: 1994; Bacelar do Nascimento / Pereira: 1996).

A programação informática realizada para este Projecto permite, actualmente, a extracção de pares de palavras com a indicação da distância a que se encontram um do outro os elementos dos pares (distâncias 1,2,3 e 4), posição que ocupam relativamente ao nó (posição anterior ou posterior) e a respectiva Frequência de ocorrência (sempre igual ou superior a 2).

O Programa Concor.cb percorre o *corpus* e determina todas as ocorrências de palavras que co-ocorrem a determinada distância (sempre $F \geq 2$). O utilizador pode indicar: a distância máxima a que pretende observar os pares e a frequência a partir da qual pretende considerar os pares. Obtém-se, pois, com este programa, uma lista de todos os pares co-ocorrentes no *corpus* com a respectiva frequência e distância a que co-ocorrem. A partir desta lista obtém-se também para cada par o conjunto especificado de pares que contêm formas pertencentes aos lemas de cada um dos elementos desse par e os contextos em que ocorrem. Elaboraram-se, ainda, programas que agrupam dados (palavras e seus pares, palavras e suas frequências) e efectuaram-se cálculos estatísticos que relacionam a frequência das palavras constituintes do par com a frequência dos pares em que ocorrem e com as frequências dessas mesmas palavras no total do *corpus*.

Apresentam-se a seguir alguns resultados obtidos, quanto a indicadores estatísticos, que permitem seriar combinações de uma forma rigorosa:

Medidas de Informação Mútua da Palavra-Nó pressão³:

Total do corpus: 9.992.706

Lema: *Pressão*

$f(x) = 1.147$

Pal. y = exercer; $f_a(x,y) = 13$; $f(y) = 325$; $f_{-2} = 5$; $f_{-1} = 8$; $f_{+1} = 0$;
 $f_{+2} = 0$; $m = -1,385$; $s^2 = 0,256$; $IC = 8,445$;

³ "O Índice Combinatório (IC) entre a palavra-nó x e a palavra co-ocorrente y é o índice calculado entre as probabilidades de ocorrência conjunta dos pares de palavras de x e de y (x, y) e de ocorrência independente das mesmas palavras. O IC é indicador do grau de significância da combinação entre as mesmas palavras.

O Índice de Fixidez (IF) da combinação obtém-se da correcção do Índice Combinatório pela consideração da variância e permite distinguir, de entre palavras com idênticos Índices de Combinatórios, aquelas em que esta é de natureza mais fixa ou menos fixa.

A variância (s^2) da distância de cada palavra da janela relativamente ao nó [...] é um bom indicador do grau de fixidez dos grupos de palavras". (Pereira 1994: 137-149).

$$\underline{IF} = 23,694$$

$$\begin{aligned} \underline{\text{Pal. y}} = \text{ceder}; \quad & \underline{fa(x,y)} = 6; \underline{f(y)} = 159; \underline{f-2} = 6; \underline{f-1} = 0; \underline{f+1} = 0; \\ & \underline{f+2} = 0; \underline{m} = -2; \underline{s^2} = 0; \underline{IC} = 8,361; \\ & \underline{IF} = 83,609 \end{aligned}$$

Realizou-se também:

- um novo programa informático que permite extrair *grupos de palavras* (ou seja palavras que ocorrem contiguamente e não com distâncias entre os seus elementos) de 5, 4, 3 e 2 unidades, com frequência igual ou superior a 2;
- uma aplicação do programa de concordâncias do Corpus de Referência do Português Contemporâneo, a este Projecto, que permite ao utilizador, ao aceder às concordâncias da palavra em estudo, saber também o tipo de texto em que ocorreu, de acordo com a tipologia estabelecida para este Dicionário, ou seja: Oral/Escrito (concordâncias sempre fornecidas separadamente) e, dentro do escrito, jornalístico, literário, técnico-científico-didáctico, político ou jurídico;
- observaram-se os resultados obtidos tendo-se realizado inúmeras análises linguísticas e experiências de aplicação com base quer na totalidade do *corpus* quer em parcelas seleccionadas dos vários tipos de discurso representados. Fizeram-se experiências tendentes a determinar quais as informações que mais interessariam aos diversos grupos de investigadores que entretanto nos foram consultando ou que considerávamos potenciais utilizadores do Dicionário.

Considerámos que os materiais a disponibilizar deveriam conter os dados exaustivos e informações extensivos sobre os mesmos, por forma a não coartar nenhuma das possibilidades de pesquisa oferecidas por este dicionário que, tendo sido inicialmente previsto como um dicionário *corpus-based*, veio a tornar-se um dicionário *corpus-driven*. Como afirmou John Sinclair, no resumo da comunicação que apresentou ao XI Encontro da Associação Portuguesa de Linguística, em Outubro de 1995, «Very few

dictionaries are *corpus-driven*, that is the lexicographers try to record without distortion the most common patterns of the language, selecting examples directly from the data, highlighting natural phraseology and pointing out regular contextual restrictions. It is argued that corpus evidence is so dramatically different from our expectations that the humbler position of the corpus-driven methodology is at present more reliable than the corpus-based one».

O produto que obtivemos pode aproximar-se do *COBUILD English Collocations on CD-ROM*, publicado em 1995 sob a direcção de John Sinclair. Esta obra parte de um inventário exaustivo como aquele que constitui o *Dicionário de Combinatórias do Português* e, na impossibilidade de incluir num CD-ROM uma tal quantidade de dados (tanto mais que o *corpus* observado era de dimensão superior ao do português), procedeu-se a uma selecção automática dos exemplos a publicar. Sendo este produto de excepcional importância, a selecção automática dos exemplos levou a que ele contenha combinatórias de menor interesse do ponto de vista quantitativo ou mesmo linguístico: «The actual selection of examples was made at random by computer with the only consideration being the space available on the CD-ROM. Because of this, you may find some unusual examples, and possibly a few offensive ones, for which we apologize. In general as we have come to expect from the Bank of English, the examples are helpful and indicate the typical phraseology of the collocation. In corpus work you quickly learn that there are some odd and unusual occurrences in natural language. Sometimes the context is not sufficiently long to explain why a word sequence has arisen, and sometimes there are mistakes, misprints, and peculiar usages. Hundreds of millions of people speak and write English all the time, and they are all different and come from all over the world, and differ in age by up to a century. The miracle is not that they don't always agree on how an English word is used, but that they agree at all» (*opus cit.* Introdução).

No intuito de comprovar a eficácia deste dicionário *corpus-driven* para determinadas aplicações, considerou-se que teria interesse apresentar, na fase final do Projecto, exemplos que resultassem das análises sistemáticas de alguns lemas, tendo-se colocado a equipa na perspectiva de utilizadores interessados, quer em seleccionar dados a integrar em dicionários gerais da língua, de formato tradicional, electrónicos ou informatizados, unilingues ou multilingues, quer em seleccionar dados para estudos de padrões combinatórios lexicais e sintácticos do português. Estas experiências de aplicação foram realizadas na fase final do dicionário, manualmente (ou seja, partindo das selecções automáticas a equipa observou e analisou dados do ponto de vista linguístico, seleccionando-os em conformidade com os resultados dessa análise).

Resultados obtidos

Apresenta-se a enumeração, acompanhada de exemplos diversificados, dos materiais disponíveis informaticamente que constituem o *Dicionário de Combinatórias do Português* e podem ser consultados no Centro de Linguística da Universidade de Lisboa.

Os utilizadores podem seleccionar todas as palavras que ocorreram em pares com Frequência igual ou superior a 2, aceder ao tipo de informação disponível sobre elas e manipular as concordâncias de acordo com os seus próprios objectivos.

Estão disponíveis, para consulta, os seguintes materiais:

a) *Corpus DCP* : 12.282.392 palavras;

b) Índice das formas lexicais (233.543 no *subcorpus* escrito e de 33.030 no *subcorpus* oral) que ocorreram no *corpus* e respectiva Frequência (F) de ocorrência, apresentadas de forma especificada conforme se trata do *corpus* de língua falada ou do *corpus* de língua escrita. (Cfr. Exemplos 1 e 2).

c) Índice dos lemas teóricos⁴ correspondentes às formas lexicais referidas na alínea b). (Cfr. Exemplos 3 e 4).

d) Índice dos 2.428.809 pares diferentes de formas lexicais que ocorreram no *corpus* com $F \geq 2$. A distância entre as formas que constituem os pares pode ser de 1 (formas contíguas), 2, 3 e 4. Sendo a palavra nó representada por A e o seu par por B, teremos:

AB / A - B / A - - B / A - - - B

BA / B - A / B - - A / B - - - A (Cfr. Exemplo 5)

Exemplo 1

Excerto da lista das formas lexicais do Corpus Escrito com a indicação da frequência das respectivas ocorrências - Intervalo: madrinha-mãezinha

<u>Forma</u>	<u>Frequência</u>	<u>Forma</u>	<u>Frequência</u>
madrinha	88	mãe-de-santo	1
madrinhas	4	mãe-filha	1
madrugada	19	mãe-ganso	1
madrugadas	587	mãe-incubadora	1
madrugador	19	mãe-natureza	4
madrugadora	7	mãe-pátria	8
	5	mãe-terra	2

⁴ Teóricos porque não sendo o *corpus DCP* um *corpus* anotado, os lemas extraídos automaticamente estão inflacionados por sobregeração, daí ser irrelevante, fornecer aqui o n.º de lemas. Contudo, a lematização automática constituiu uma fase interessante do Projecto pois permitiu estabelecer índices associativos que têm como *nó* um vocábulo e todas as suas formas flexionadas.

madrugadoras	2	mãe-tia	1
madrugadores	3	mãe	3442
madrugar	5	maeght	2
madrugava	1	maelstrom	1
madrugavam	1	maertens	2
madrugueiros	1	mães-modelo	1
madura	53	mães	349
maduramente	1	maestra	4
maduras	13	maestranza	1
madureira	22	maestras	1
madureiro	1	maestria	2
madurez	1	maestriño	1
madureza	2	maestro	154
maduro	58	maestros	9
maduros	44	mãezinha	81

Exemplo 2

Excerto da lista das formas lexicais do Corpus Oral com a indicação da frequência das respectivas ocorrências

Intervalo: madrinha-mãezinha.

<u>Forma</u>	<u>Frequência</u>	<u>Forma</u>	<u>Frequência</u>
madrinha	28	maduro	1
madrinhas	1	mãe	385
madrugada	7	mães	25
madrugadas	2	maestro	3
maduras	3	mãezinha	15

Exemplo 3

Excerto do índice dos lemas teóricos correspondentes às formas lexicais (Corpus Escrito) referidas no exemplo anterior

Forma (Lema)

madrinha (madrinha madre)

madrinhas (madrinha madre)
madruga (madrugar)
madrugada (madrugada madrugar)
madrugadas (madrugada madrugar)
madrugador (madrugador)
madrugadora (madrugador)
madrugadoras (madrugador)
madrugadores (madrugador)
madrugar (madrugar)
madrugava (madrugar)
madrugavam (madrugar)
madruguemos (madrugar)
madura (maduro madurar)
maduramente (maduramente)
maduras (maduro madurar)
madureira (madureira)
madureiro (madureiro)
madurez (madureza)
madureza (madureza)
maduro (maduro madurar)
maduros (maduro)
mãe-de-santo (mãe-de-santo mãe santo)
mãe-filha (mãe filho)
mãe-ganso (mãe-ganso mãe ganso)
mãe-incubadora (mãe-incubadora mãe incubadora)
mãe-pátria (mãe-pátria mãe pátria)
mãe-terra (mãe-terra mãe terra)
mãe-tia (mãe-tia mãe tia)
mãe (mãe)
maeght (NULO)
maelstrom (NULO)
maertens (NULO)
mães-modelo (mãe-modelo mãe modelo)
mães (mãe)
maestra (NULO)
maestranza (NULO)

mastras (NULO)
maestria (maestria)
maestriño (NULO)
maestro (maestro)
maestros (maestro)
mãezinha (mãe)

Exemplo 4

Excerto do índice dos lemas teóricos correspondentes às formas lexicais (*Corpus Oral*) referidas no exemplo anterior

Forma (Lema)

madrinha (madrinha madre)
madrinhas (madrinha madre)
madrugada (madrugada madrugar)
madrugadas (madrugada madrugar)
maduras (maduro madurar)
maduro (maduro)
mãe (mãe)
mães (mãe)
maestro (maestro)
mãezinha (mãe)

Exemplo 5

Dim	Fre	Dist	Pos1	Pos2	Pos3	Pos4	Pos5
2	2	1	abelha	e			
2	7	1	abelha	na			
2	7	1	abelhas	operárias			
2	3	1	abelhas	que			
2	6	1	nova	abelha			
2	2	1	rainha	abelha			
2	2	2	geração	abelhas			
2	2	2	jornal		abelha		
2	7	2	abelha		chuva		
2	2	2	abelhas		e		

2	2	2	abelhas	nascidas
2	2	3	abelha	a
2	3	3	abelha	da
2	2	4	abelha	é
2	3	4	abelhas	a
2	3	4	abelhas	flores
2	2	4	se	abelhas

Legenda:

Dim - Dimensão do grupo (neste caso estamos sempre a tratar de pares de palavras).

Fre - Número de vezes que o par ocorreu no *Corpus*.

Dist - Distância entre as formas que constituem o par.

Posi - Posicionamento da forma no par, com variação possível de 1 a 5

Destes índices constam:

- Frequência de ocorrência da palavra-nó no *corpus*
- Frequência de ocorrência da palavra-nó no par, ou seja, a frequência do par
- O Índice Combinatório (IC) do par (cfr. nota 2) (Cfr. Exemplo 6).

Exemplo 6

Excerto dos índices combinatórios dos pares do lema Célebre

Frequência no corpus escrito de <u>Célebre</u> :		454
Par	<u>Célebre - Tristemente</u> :	<u>F</u> = 10
Índice Combinatório:		<u>IC</u>
	tristemente célebres	8,825
	tristemente célebre	8,480
Par	<u>Célebre - Ficar</u> :	<u>F</u> = 14
	ficaram célebres	6,203
	ficou célebre	4,836
Par	<u>Célebre - Frase</u> :	<u>F</u> = 7

e) Índices de lemas co-ocorrentes (Cfr. Exemplo 7).

Exemplo 7

Lista de lemas co-ocorrentes de Célebre

Escrito

FT = 454; FC = 1.360

10	Tristemente	(real:10)
4	Criminoso	(real:4)
14	Ficar	(real:14)
7	Frase	(real:7)
12	Tornar	(real:12)
6	Autor	(real:8)
10	Muito	(real:10)
325	De	(real:422)
37	Mais	(real:37)
45	Ser	(real:47)
10	Tão	(real:10)
11	Sua	(real:20)
73	Em	(real:96)
6	Seus	(real:6)
31	Por	(real:36)
4	Dia	(real:4)
6	Já	(real:6)
42	Que	(real:67)
46	E	(real:77)
14	Com	(real:18)

f) Listas brutas de concordâncias de todos os pares com 1 linha de contexto,

contendo toda a informação quantitativa e estatística relativamente ao lema e às formas e a identificação do tipo de discurso a que pertence, de acordo com a tipologia estabelecida e enunciada na alínea a) (Cfr. Exemplo 8).

Exemplo 8

Lista bruta de concordâncias de todos os pares de Amplo com 1 linha de contexto

*** Escrito ***

*** FT 483 Amplo ***

*** FC 604 Amplo ***

*** 4 Abranger (real:4) ***

amplos abrangendo 8,773

2 amplos abrangendo 1

é concebido em termos amplos, abrangendo todos os

concebido em termos amplos, abrangendo todos os

amplo abrangendo 7,571

2 amplo abrangendo 1

está empregado em sentido amplo, abrangendo as pr

empregado em sentido amplo, abrangendo as

*** 17 Consenso (real:17) ***

amplos consensos 8,773

2 amplos consensos 1

no PS estão em moda os amplos consensos. Até às

no Pontal gerar amplos consensos fora das

amplo consenso 7,856

15 amplo consenso 1

em promover o mais amplo consenso democrático para estabelecer um amplo consenso em volta desta resultar de um amplo consenso entre todas as fo Resultado de amplo consenso estabelecido se houvesse um amplo consenso mas afinal basta encontrada na base de amplo consenso nacional, o menos sem um amplo consenso, ou na Assembleia ter sido resultado de amplo consenso político, defender «o mais amplo consenso possível» entre para se encontrar o mais amplo consenso que se e depois de obtido um amplo consenso. Um piscar

As concordâncias aparecem indexadas por ordem decrescente dos Índices Combinatórios (também é possível indexá-las com outras ordenações). Dentro de cada conjunto de concordâncias de um par, os contextos podem ser ordenados por ordem alfabética das palavras que vêm à direita ou à esquerda do par ou, simplesmente pela ordem de ocorrência no *corpus*.

Optámos por não classificar morfossintacticamente os lemas para não impor limitações às possibilidades de análise que o Dicionário oferece. Por exemplo nos grupos nominais, constituiria uma limitação para o utilizador a separação dos contextos segundo a sua realização como substantivos ou como adjectivos, ou ainda, no caso dos Particípios Passados, será certamente muito produtivo que o próprio utilizador possa estabelecer os seus critérios quanto à realização adjectival, de acordo com a observação de todas as ocorrências.

g) Listas de concordâncias com contexto expandido até à dimensão adequada aos objectivos do utilizador.

h) Listas de grupos contíguos de 5, 4, 3 e 2 palavras com as respectivas Frequências de ocorrência e contextos em que ocorreram (Cfr. Exemplo 9).

Exemplo 9

Excerto da lista de concordâncias dos grupos de 3 palavras cujo nó é Amplo

ente educativa e tornando-o num amplo espaço de realização do cidadão, p
que "não sejamos somente "um amplo espaço de comércio livre" e tenha
sicamente, pela criação de um amplo espaço de lazer, de convívio, pedonal
marcha, já há algum tempo, um amplo movimento de carácter independen
uma forma de "desmobilizar" um amplo movimento de contestação à nov
esta guerra foi o início de um amplo movimento de descolonização: várias
r o desemprego, foi traçado um amplo programa de construção de obras p
Bravo, que participará de um amplo programa de provas em Portugal e no
r o país. Aristide anunciou um amplo programa de reconciliação nacional,
pes, 25, ao Saldanha, possui ampla área de serviços com 280 metros quadr
enovadas da população, é a ampla área de que dispõe para se estender. O pe
io -, o diploma introduz uma ampla cláusula de exclusão da sanção, cuja g
io -, o diploma introduz uma ampla cláusula de exclusão da sanção, cuja g
ndo as pernas perras na saia ampla e comprida, levou-me até à sala e apont
ha cor de ervilha seca, a saia ampla e comprida, o casaquinho curto muito j
is e na base de uma reflexão ampla e não em termos especiais para os tribu
via ter-se agido com visão ampla e não ao sabor dos calendários eleitorais
ões básicas para desenvolver uma ampla oferta turística em torno do impor
ões básicas para desenvolver uma ampla oferta turística em torno do impor
calma, uma "discussão pública ampla profunda e cautelosa" , pois "essa di
calma, uma "discussão pública ampla profunda e cautelosa" , pois "essa di
em antes ter passado por uma ampla reflexão no seio do executivo munic
em antes ter passado por uma ampla reflexão no seio do executivo munic
ão hipóteses de trabalho mais amplas e com outras contrapartidas. No amo
ão hipóteses de trabalho mais amplas e com outras contrapartidas. Aguarde
to em seda, sobre umas calças amplas em crepe e cabeção em organza pliss
to em seda, sobre umas calças amplas em crepe e cabeção em organza pliss
963 está empregado em sentido amplo, abrangendo as prestações pecuniár
963, está empregado em sentido amplo, abrangendo as prestações pecuniár
ra garantir "um cada vez mais amplo acesso dos cidadãos à ópera. A polític

que garanta um cada vez mais amplo acesso dos cidadãos à ópera». Ter m
um carro da polícia no seu mais amplo campo de visão. Aos sábados e do

i) Resultados de análises linguísticas realizadas manualmente sobre 26.568 pares diferentes (correspondentes a 60 nós) com o objectivo de demonstrar o interesse da reutilização deste dicionário de base em aplicações lexicográficas quer de carácter mais geral (selecção visando a delimitação de grupos lexicalizados ou com tendência para a lexicalização) quer de carácter mais especializado (selecção tendo em vista, por exemplo, dicionários de tipo analógico (co-ocorrências descontínuas) ou sintáctico (co-ocorrências que apontam para a selecção de estruturas sintácticas) (Cfr. Exemplo 10).

Exemplo 10

Excerto do resultado da análise linguística realizada manualmente para Abaixo

*** *Escrito* ***

*** *FT 877* Abaixo ***

*** *FC 3692* Abaixo ***

*** *7* Negociar (real:7) ***

negociando-se abaixo 8,574

2 negociando-se abaixo 1

as alemãs. O dólar foi atingido, negociando-se abaixo dos 1, 52 m

mana enfraquecido face ao marco, negociando-se abaixo dos 1, 52 m

negociar-se abaixo 8,391

5 negociar-se abaixo 1

ores prevêem que o marco volte a negociar-se abaixo dos 103 escud

Travessa O marco voltou ontem a negociar-se abaixo dos 103 escud

ESCUDOS O marco voltou ontem a negociar-se abaixo dos 103 escud

passada 25 por cento, chegando a negociar-se abaixo dos dois mil

passada 25 por cento, chegando a negociar-se abaixo dos dois mil

*** 4 Sumir (real:4) ***

sumido abaixo 7,293

2 sumido abaixo 3

ceber... esse mistério de me ter sumido pelo chão abaixo.» «Ah! a
lado, como se o herói se tivesse sumido pelo chão abaixo. Ficou i

sumir-se abaixo 7,092

2 sumir-se abaixo 3

essoa. Alagado em suor, ansiando sumir-se pelo chão abaixo, encol
suores frios, e o seu desejo era sumir-se pelo chão abaixo. Mas,

*** 25 Deitar (real:25) ***

deitaram abaixo 6,951

3 deitaram abaixo 1

Surgiram os impugnadores e deitaram abaixo a arce maravilho
e vermelhas. No local onde deitaram abaixo a praça de touro
da Avenida da Liberdade, deitaram abaixo a velha praça, a

deitam abaixo 6,399

2 deitam abaixo 1

pequenos remoques que nos deitam abaixo São pequenas frase
pequenos remoques que nos deitam abaixo E AINDA 10 Elas dã

deitados abaixo 6,376

2 deitados abaixo 1

qualquer Ordem, com livros deitados abaixo para a consulta,
Havia lances de passeios deitados abaixo, com brechas pro

deitou abaixo 6,281

4 deitou abaixo 1

se soltou. Sorria triunfante, e deitou abaixo as calças do pijam
edilidade mais corajosa a deitou abaixo para alargar o loc
pelo PS, de uma só penada, deitou abaixo, para já do ponto
de Oleiros. O município deitou abaixo recentemente algun

deitar abaixo 5,783

4 deitar abaixo 1

aí alguma revolução para o deitar abaixo - acudiu D. Doroté
adianta: "Se decidíssemos deitar abaixo alguma destas casa
e o dever para mandar deitar ABAIXO o centro cultural
dia após dia; quando vemos deitar abaixo ou deixar cair vel

3 deitar abaixo 3

satisfeita. Não vale a pena deitar a bainya abaixo nem consi
mandara quebrar os mastros, deitar as amuradas abaixo, rasga
pequena se fica aí capaz de deitar sete casas abaixo! Porém

deita abaixo 5,650

2 deita abaixo 1

vez construídos, ninguém os deita abaixo. Como resultado de
conhecer (mas já presentes) deita abaixo gente como eu, caíd

deitada abaixo 5,539

2 deitada abaixo 1

partidos, uma amurada deitada abaixo e as velas em far
segurança? A rede está toda deitada abaixo.» Este é, visível

deitado abaixo 5,435

Venables, demite-te") foi deitado abaixo pelos claros 4-1
Oleiros. O município havia deitado abaixo recentemente algu
desde que soube que tinham deitado abaixo a casa do ervanár

Como consta da apresentação da candidatura do Projecto, não foi prevista, no seu âmbito, a publicação dos Resultados. Considera-se, no entanto, que a enorme quantidade de informação que está disponível (para consulta) e as reutilizações que permite deveriam constituir matéria para Projectos exclusivamente destinados a publicações. São, desde já, previsíveis, a curto prazo, a publicação em CD-ROM, de excertos seleccionados dos dados brutos (torna-se impossível incluir num CD-ROM a totalidade dos resultados) assim como a publicação em suporte informático e em papel, de ocorrências lexicais restritas para aplicação em dicionários gerais (monolingues ou bilingues) e de co-ocorrências lexicais e gramaticais para aplicações pedagógicas. A longo prazo, o Projecto configura-se como fonte inesgotável de estudos e, conseqüentemente, de Publicações, as mais diversas.

Conclusões

A investigação sobre um *corpus* permitiu a observação controlada de grandes massas de dados, de forma a determinar tendências fortes no uso da língua, factor fundamental em trabalhos de linguística teórica, descritiva e aplicada e em trabalhos interdisciplinares, designadamente em linguística computacional, estatística lexical, psicolinguística e sociolinguística.

A investigação sobre *corpora* tem tornado patente que as línguas naturais são, quer do ponto de vista paradigmático, quer do ponto de vista

sintagmático, mais padronizadas do que os estudos baseados na intuição faziam prever. É também o estudo de *corpora* que permite identificar esses padrões (léxico-sintático-semântico) e a regularidade com que ocorrem e interagem nos diversos usos.

No caso do *Dicionário de Combinatórias do Português*, os fenómenos associativos contínuos e descontínuos que determinámos parecem ser de natureza diversa: textual (observada repetidamente em determinados textos ou tipos de textos, potencializam análises textuais e estilísticas), discursiva (observados em situações enunciativas diversas, terão papel importante na análise do discurso) e, em grande medida, cognitiva (serão fundamentais em estudos sobre a aquisição e a memória lexical dos indivíduos); estes fenómenos representam, pois, relações diversas que os utilizadores podem analisar em termos quantitativos e qualitativos, utilizando as ferramentas informáticas disponíveis, para identificar a extensão e os tipos particulares de relação associativa. Não se trata de fornecer restrições de selecção das associações baseadas em conceitos gramaticais ou de congruência semântica ou, ainda, de natureza pragmática, noções essas intuitivas e que supõem a correlação permitido/*vs*/proibido. O que está à disposição dos utilizadores são selecções preferenciais observáveis e procedimentos informáticos e estatísticos que os ajudarão a distinguir factores relevantes do puro ruído e também a isolar fenómenos ou induzir generalizações.

Bibliografia

Bacelar do Nascimento, Maria Fernanda / Pereira, Luísa Alice Santos (1996): «Dicionário de Combinatórias do Português: associações lexicais frequentes observadas num corpus de português contemporâneo», em: Faria, Isabel Hub / Correia, Margarita (eds.) (1996): *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística* -

Dicionários, Lisboa: APL, Volume II, págs. 43-54.

Benson / Benson / Ilson (eds.) (1986): *The BBI Combinatory Dictionary of English: A guide to Word Combinations*, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Biber (1993): «*Co-ocurrence Patterns among Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition*», em: *Computational Linguistics* 19-3, págs. 531-538.

Calzolari, Nicoletta / Bindi (1990): «Acquisition of lexical information from a large textual italian corpus», em: Karlgreen, Hans (eds.), *Coling 90*, Papers presented to the 13th International Conference on Computational Linguistics, Hensinky: Hensinky University.

Church / Hanks (1990): «Word association norms, mutual information, and lexicography», em: *Computational Linguistics* 16 (1), págs. 22-29.

Firth (1955): «Modes of meaning», em: *Papers in Linguistics 1934-1951*, págs. 190-215.

Pereira, Luísa Alice Santos (1994): *Como se combinam as palavras? Contributo para um Dicionário de Combinatórias do Português*, (Dissertação de Mestrado), Lisboa: Faculdade de Letras da Universidade de Lisboa.

Sinclair, John (1991a): *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Sinclair, John (1991b): «The automatic analysis of corpora», em: *Interim Report - December 1991 of the EC Project Network of European Reference Corpora*, Pisa: Istituto di Linguistica Computazionale.

Sinclair, John (1996): «Tipologia Textual EAGLES», em: Bacelar do Nascimento, Maria Fernanda / Rodrigues, Maria Celeste / Gonçalves, José Bettencourt (eds.) (1996): *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística - Corpora*, Lisboa: APL, Volume I, págs. 39-91.

Smadja (1993): «Retrieving Collocation from Text: Xtract», em: *Computational Linguistics*, 19 (1), págs. 143-177.