

Um novo léxico de frequências do português

Maria Fernanda Bacelar do Nascimento

Centro de Linguística da Universidade de Lisboa

Um novo léxico de frequências do português, recentemente estabelecido a partir de um *corpus* de grandes dimensões, veio colmatar uma lacuna que de há muito se fazia sentir nos estudos sobre o português contemporâneo. De facto, os léxicos de frequência de que dispúnhamos baseavam-se em *corpora* de pequenas dimensões e pouco diversificados na sua constituição interna. É o caso do vocabulário do Português Fundamental cujas 2.217 entradas foram extraídas, essencialmente, de um *corpus* oral de 700.000 palavras construído na década de 70¹ (*Português Fundamental, Vocabulário*, 1987) e do *Frequency Dictionary of Portuguese Words* estabelecido com base num *corpus* literário de 500.000 palavras (Duncan, 1972).

Os estudos baseados em grandes *corpora* desenvolveram-se extraordinariamente nas últimas décadas e têm demonstrado a excelência destes recursos linguísticos numa enorme variedade de áreas de investigação teórica e aplicada, sendo cada vez mais utilizados na identificação e interpretação de aspectos quantitativos e probabilísticos das línguas e em análises qualitativas que têm como objectivo a compreensão de factos linguísticos, empiricamente observados nos reais e muito diversos contextos de uso em que ocorrem.

No que respeita aos estudos do léxico, a importância atribuída à dimensão dos *corpora* já era reconhecida por Zipf que, em 1935, baseou os seus princípios de distribuição das frequências lexicais em amostragens de textos com uma extensão muito significativa (Zipf, 1935). Hoje, a possibilidade que temos de analisar automaticamente *corpora* de muitos milhões de palavras confere grande fiabilidade aos resultados obtidos, mesmo no que respeita às frequências menos altas. Como reconhece M.A.K. Halliday, «if we want to investigate any

¹ Para uma informação detalhada acerca do estabelecimento do vocabulário do Português Fundamental, cf. Bacelar do Nascimento, *et al.*, 1987 e Bacelar do Nascimento, *et al.*, 1987.

words other than those of highest frequency in a language, we need text data running at least into millions of words, and preferably into hundreds of millions» (Halliday, 1993: 1).

No âmbito do Projecto "Léxico multifuncional computadorizado do português contemporâneo",² concluiu-se em Dezembro de 2000 um *Léxico de frequências* de 26.980 vocábulos (entradas lexicais) e das 140.976 formas lematizadas desses vocábulos, extraído de um *corpus* de 16.210.438 palavras³ do português europeu. As entradas lexicais que constituem este Léxico atingiram, no *corpus*, frequências iguais ou superiores a 6 e cada uma dessas entradas é seguida de informação gramatical (categoria morfossintáctica) e de informação quantitativa (nível de ocorrência no *corpus*); as mesmas informações são dadas para todas as formas flexionadas e compostas de cada vocábulo; para estas formas dá-se ainda informação sobre o tipo de *sub-corpus* - oral/escrito - em que ocorreram.

O corpus

Para a realização do projecto, foi desenhado e extraído do *Corpus de Referência do Português Contemporâneo* (CRPC) do CLUL⁴ um *corpus* de 16.210.438 palavras - o CORLEX, que contém um *sub-corpus* de língua escrita (15.354.243 palavras) e um *sub-corpus* de língua falada (856.195 palavras).

O CORLEX cumpre os princípios comumente estabelecidos e as recomendações internacionais sobre a dimensão e o desenho de *corpora* linguísticos, de carácter geral, que se destinem à extracção de léxicos (por exemplo, Zampolli, 1995). De facto, para estudos desta natureza, consideram-se aceitáveis *corpora* com dimensões entre 10 e 20 milhões de palavras. O *corpus* original (the Main Corpus) que esteve na base do primeiro *Collins Cobuild English*

² Este projecto, subsidiado pelo programa PRAXIS XXI (PRAXIS XXI/2/2.1/CSH/759/95), iniciou-se em 1997 e terminou em 2000. Foi realizado, em parceria, pelo Centro de Linguística da Universidade de Lisboa (CLUL) - instituição coordenadora do projecto - , pelo Instituto de Engenharia de Sistemas e Computadores (INESC) e pela Editorial Verbo.

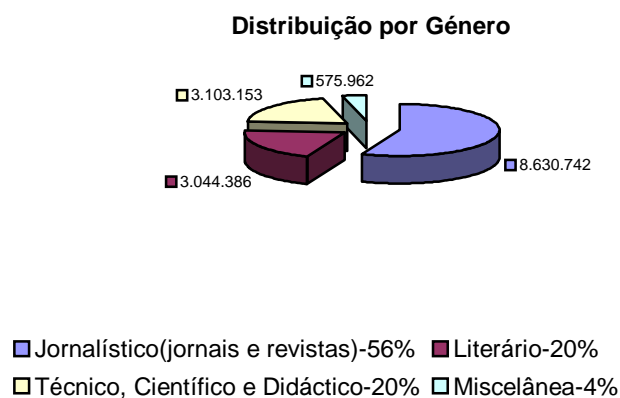
³ Em todos os casos em que refiro a dimensão do *corpus*, *palavra* é sinónimo de *ocorrência*.

⁴ *Corpus* aberto em contínuo desenvolvimento. Este *corpus* monitor, à data da conclusão do Léxico (Dezembro 2000), continha 150 milhões de palavras.

Language Dictionary elaborado pelo grupo de investigação do Departamento de Inglês da Universidade de Birmingham continha 7 milhões de palavras (Sinclair, 1995: vii). Quanto ao desenho do *corpus*, também se considera que a estrutura interna do CORLEX é adequada aos fins a que se destina. Dela fazem parte textos orais e escritos que cobrem uma grande variedade de tipos de linguagem, sendo a diversidade de géneros e de matérias uma dominante deste *corpus*.⁵

Constituição do sub-*corpus* escrito (15.354.243 palavras)⁶

Deste *sub-corpus* fazem parte textos de diversos géneros. A maior extensão do *corpus* jornalístico (56% do *corpus* escrito e 53% do *corpus* total) teve em vista dar predomínio a uma linguagem comum, isto é, não especializada, e cobertura a uma grande variedade de temas. No gráfico "Distribuição por Género" pode observar-se a constituição, nesse domínio, do *sub-corpus* escrito.



As recolhas foram feitas em diversas *Fontes*. De um modo geral, e sempre quando se trata do género literário, o *corpus* é constituído por *amostragens* dos títulos seleccionados.

⁵ O CORLEX é constituído por ficheiros nos formatos Text Only, HTML e SGML.

Fontes do <i>sub-corpora</i> jornalístico			
Jornais			
Títulos dos jornais	Datas	Nº de exemplares	Nº de artigos
Diário de Notícias (DN)	1997 e 1998	21	1.963
Jornal de Notícias (JN)	1997 e 1998	47	6.425
Público (PU)	1997 e 1998	37	4.697
Revistas			
Título das revistas	Datas	Nº de exemplares	Nº de artigos
Máxima	1992 a 1994	13	986
Grande Reportagem	1992 a 1995	21	832
Visão	1996 e 1997	18	1820

Fontes do <i>sub-corpora</i> literário (Romances, Novelas, Contos, Poesias, Memórias e Peças de teatro de autores portugueses)		
Nº de Autores	Nº de Títulos	Datas
135	186	séc. XIX (2ª metade): 11 autores; 14 títulos
		séc. XX: 124 autores; 172 títulos

Fontes do <i>sub-corpora</i> Científico, Técnico e Didáctico⁷		
Nº de Autores ⁸	Nº de Títulos	Datas
91: livro científico e técnico - 68 livro didáctico - 23	93: livro científico e técnico - 68 livro didáctico - 25	1980 - 1993

Fontes do <i>sub-corpora</i> "Miscelânea"		
Tipo de documento	Nº de textos/artigos	Datas
Jornais e revistas especializados	347	1900 - 1997
Outros documentos	30	

⁶ Uma parte deste *sub-corpora* é constituída por materiais cedidos pela Editorial Verbo.

⁷ Níveis de ensino a que se reportam os livros didácticos: 5º a 11º ano de escolaridade.

⁸ Autorias colectivas foram contabilizadas como um só autor.

Constituição do *sub-corpus* oral (856.195 palavras)

O *sub-corpus* oral é constituído pela transcrição ortográfica do registo magnético de conversas informais e de produções mais formais (conferências, entrevistas na rádio e na televisão, etc.).

<i>Sub-corpus</i> Oral			
Tipo de discurso	Nº de palavras	Nº de textos	Datas
espontâneo	752.394	1409	Décadas de 1970 e 1990
formal	103.801	150	Década de 1980

O Léxico

A extracção do léxico

A primeira operação de extracção do léxico consistiu na indexação de todas as formas lexicais diferentes que ocorreram no CORLEX. Observou-se, então, que em 16.210.438 ocorrências se registaram 283.530 formas diferentes. A forma que atinge maior Frequência é a forma "de" e 128.383 formas têm Frequência 1.

Todas estas formas foram automaticamente etiquetadas (etiquetagem morfossintáctica) e lematizadas;⁹ as etiquetas atribuídas a cada forma dizem respeito às categorias morfossintácticas teoricamente atribuíveis a essas formas. Tendo em conta pois todas as possibilidades de categorização de cada forma, a lematização feita automaticamente originou 39.696 lemas (vocábulo).

Seguidamente, procedeu-se a uma verificação da lematização e das etiquetas atribuídas a todas as formas que integravam lemas com uma frequência de ocorrência teórica

⁹ Nestas operações foi utilizado o anotador automático PALAVROSO do INESC; esta ferramenta é usada também no CLUL, tendo sido cedida a esta instituição por permuta com um *corpus* de treino, ao abrigo de um protocolo de intercâmbio estabelecido em 6 de Janeiro de 1992 entre as duas instituições.

superior a 6.¹⁰ Nesta verificação seguiram-se os critérios de classificação e de lematização utilizados no projecto Português Fundamental (Bacelar do Nascimento *et al.*,1987: 358-391). Esta verificação manual resultou em correcções de vários tipos:

- Inclusão de casos de homografia não considerados pelo anotador automático.

Exemplos:

Forma	Categorias consideradas pelo anotador automático	Categorias consideradas depois da verificação manual
apoiante	Adjectivo	Adjectivo e Nome
pelarias	Verbo	Verbo e Nome

- Anulação de lemas que eram fruto de sobregeração efectuada pelo anotador automático.

Exemplos:

aromar (verbo)	ciumar (verbo)
aurorar (verbo)	fortalezar (verbo)
bruxar (verbo)	fritir (verbo)
cavaleirar (verbo)	

- Afastamento de casos de homografia de realização muito improvável no CORLEX, face aos dados observados.

¹⁰ Esta verificação foi feita manualmente pela equipa do CLUL.

Exemplos:

Forma ¹¹	Categorias consideradas pelo anotador automático	Categorias previsíveis no CORLEX
adega	Nome e Verbo	Nome
adegas	Nome e Verbo	Nome
alameda	Nome e Verbo	Nome
alamedas	Nome e Verbo	Nome

- Análise e classificação de formas não reconhecidas (classificação: nulo) pelo anotador automático, tais como certas siglas, acrónimos, estrangeirismos, abreviaturas, formas hifenadas, formas cuja ortografia não corresponde à ortografia actualmente em vigor e, ainda, advérbios terminados em *-mente*, diminutivos em *-inho* e *-ito* e outras.

Exemplos:

Exemplos de:	Forma	Frequência ¹²
Siglas	bd	64
	irs	256
Acrónimos	frelimo	63
	palop	61
Estrangeirismos	crochet	36
	homepage	488
	motard	22
Abreviaturas	dra	50
	sra	28
Formas hifenadas	anos-luz	30
	ex-director	37
	há-de	844
	pára-quadristas	38
Ortografias desviantes	êle	569
	sêde	24
Advérbios em <i>-mente</i>	comprovadamente	11
	merecidamente	10

¹¹ As únicas formas dos lemas ADEGAR e ALAMEDAR eram adega/adegas e alameda/alamedas.

¹² São apresentadas as frequências destes exemplos "nulos" para que se possa ter uma ideia dos dados recuperados por este processo e, conseqüentemente, dos benefícios para os anotadores automáticos da sua aplicação a *corpora*. A título de exemplo, registre-se que foram classificadas como "nulos" pelo anotador automático 5.967 formas em *-inho* e 3.277 formas em *-ito*, posteriormente recuperadas manualmente.

Exemplos de:	Forma	Frequência
Outros	clonagem	102
	implosão	99
	metadona	41
	pedófilo	60
	racionalismo	52

Feitas estas verificações, procedeu-se a nova lematização teórica¹³ que veio a dar os seguintes resultados:

Número de lemas com Frequência superior a 6 _____ 30.806

Número de formas diferentes _____ 131.433

Número de formas homógrafas _____ 44.773

Desambiguação

Para a desambiguação das formas homógrafas seguiram-se procedimentos automáticos¹⁴ e manuais. Dada a grande extensão do *corpus*, os níveis de frequência a fornecer foram calculados probabilisticamente (excepto nos casos em que se procedeu à desambiguação manual da totalidade das ocorrências). Como base para cálculos estatísticos, cálculos de probabilidades e extração automática de regras foi utilizado o *sub-corpus* português anotado PAROLE.¹⁵

¹³ Esta lematização teórica serviu de base a uma aferição destes dados empíricos relativamente a léxicos construídos sem recurso a *corpora*. Este trabalho foi realizado pela Editorial Verbo e os resultados da aferição constam do Relatório Final do projecto.

¹⁴ *Software* do INESC usado nesta fase do trabalho: Sistema Interactivo de Desambiguação de *Corpora* DESAMBIG e ferramentas integradas no ENCONTRA&ESTATIC.

¹⁵ *Sub-corpus* de 250.000 formas anotadas que faz parte do *corpus* português do projecto *Preparatory Action for Linguistic Resources Organisation for Language Engineering* (PAROLE) da Comissão das Comunidades Europeias em que participaram 18 instituições europeias (Cfr. internet: <http://www.linglink.lu/le/projects/le-parole>). Neste projecto, o *corpus* e o léxico portugueses foram estabelecidos e anotados pelo CLUL - instituição coordenadora da parte portuguesa - e pelo INESC. O *sub-corpus* aqui referido foi anotado com o anotador morfológico PALAVROSO e desambiguado (em conformidade com critérios comuns estabelecidos para todas as línguas representadas no PAROLE) por uma equipa CLUL/INESC que utilizou o DESAMBIG como ferramenta auxiliar da análise manual das formas em contexto.

Desambiguação automática

De acordo com instruções dadas pelo INESC, procedeu-se, no CLUL, a uma preparação do CORLEX, que consistiu em retirar e converter as marcas de HTML e de SGML para Text Only; retirar as codificações estabelecidas internamente pelo CLUL; segmentar todo o *corpus* em "frases", isto é, introduzir a indicação "fim de linha" após cada sinal de pontuação forte.

Terminada a preparação do *corpus*, o INESC adaptou e correu sobre o CORLEX o desambiguador *Eric Brill's Tagger*.¹⁶

Amostragem do CORLEX desambiguado automaticamente:

```
-/O A/T Câmara/N Municipal/A de/S Portimão/N abriu/V concurso/N  
para/S a/T construção/N de/S um/T edifício/N de/S oito/M pisos/N  
num/S total/N de/S 16/M fogos/N habitacionais/A ,/O sendo/V 5/M  
do/S tipo/N T-1/M ,/O 5/M do/S tipo/N T2/N e/C seis/M do/S tipo/N  
T3/M ,/O mais/R dois/M espaços/N destinados/V a/T comércio/N ,/O  
indústria/N ou/C serviços/N no/S piso/N térreo/A ./O  
Não/R foi/V indicada/V base/N de/S preço/N ,/O que/P deve/V ser/V  
apresentada/V pelos/S concorrentes/N juntamente/R com/S os/T  
respectivos/A projectos/N ,/O bem/R como/C o/T prazo/N de/S  
construção/N ./O
```

Desambiguação manual¹⁷

Observado o *sub-corpus* anotado PAROLE (250.000 formas), verificou-se existirem algumas diferenças entre a categorização do PAROLE e a do CORLEX, nomeadamente:

a) certas formas (ambíguas) que ocorreram no CORLEX não ocorreram no PAROLE (cf. exemplos "capricho" e "vindima" no quadro abaixo);

Parte dos recursos linguísticos portugueses construídos no âmbito do PAROLE está disponível e é distribuída pela European Language Resources Association (ELRA). Referências destes recursos linguísticos no catálogo ELRA: ELRA-W0024/01 PAROLE Portuguese *Corpus*; ELRA-W0024/02 PAROLE Portuguese *Sub-Corpus*; ELRA-L0035 PAROLE Portuguese Lexicon.

¹⁶ Acessível a partir de <http://www.cs.jhu.edu/~brill> ou ftp://ftp.cs.jhu.edu/pub/brill/Programs/RULE_BASED_TAGGER_V.1.14.tar.Z.

¹⁷ As desambiguações manuais foram realizadas pela equipa do CLUL.

b) certas formas foram apenas classificadas no PAROLE com algumas das categorias morfossintáticas com que foram categorizadas no CORLEX (cf., por exemplo, "fora" no quadro abaixo);

c) algumas formas foram classificadas no PAROLE com as mesmas categorias atribuídas no CORLEX mas, naquele projecto, as formas não foram integradas em lemas distintos (cf. exemplos "revista" e "visto" no quadro abaixo).

Assim, procedeu-se à análise manual dos contextos dessas formas, quer por amostragens - quando a Frequência de ocorrências dessas formas era muito alta - quer na sua totalidade - quando a Frequência de ocorrência era igual ou inferior a 200.

Exemplos:

Formas	Categorias existentes no <i>corpus</i> anotado PAROLE	Categorias previsíveis no CORLEX	Frequência no CORLEX
acrescente	—	Adjectivo; Nome; Verbo	79
capricho	—	Nome; Verbo	105
saliente	—	Adjectivo; Verbo	82
vindima	—	Nome; Verbo	71
fora	Advérbio; Verbo	Advérbio; Elemento de locução; Interjeição; Nome; Preposição; Verbo (ser); Verbo (ir)	6.595
revista	Nome; Verbo	Nome; Verbo (rever); Verbo (revistar)	1.152
visto	Adjectivo; Nome; Verbo	Adjectivo; Nome; Verbo (ver); Verbo (vestir)	1.618

Outras análises manuais realizadas nesta fase do projecto:

- Observação de formas participiais cuja classificação automática como Verbo e Adjectivo suscitou dúvidas.

Exemplos:

Formas	Frequência	Categorias consideradas pelo anotador automático	Resultados da desambiguação manual	Frequência
poisada	12	Adjectivo e Verbo	Verbo	12
encantadas	1	Adjectivo e Verbo	Adjectivo	1
encantado	9	Adjectivo e Verbo	Adjectivo	9
encantados	6	Adjectivo e Verbo	Adjectivo	6

- Desambiguação de formas com a mesma categoria gramatical mas pertencentes a lemas diferentes.

Exemplos:

Formas	Categoria	Lemas diferentes a que pertencem
consumo	Verbo	consumir
		consumar

- Pesquisa de formas multilexicais não hifenadas a partir de formas hifenadas correspondentes.

Exemplos:

Formas hifenadas	Frequência	Formas correspondentes não hifenadas	Frequência
quarto-de-hora	7	quarto de hora	85
rés-do-chão	161	rés do chão	3
não-instrumental	6	não instrumental	5
cor-de-laranja	8	cor de laranja	8

- Verificação de formas que pela sua alta Frequência no *corpus* sugeriam a hipótese de que, em grande parte, tivessem ocorrido como Nomes Próprios.

Exemplos:

Forma ¹⁸	Frequência	Classificação dada pelo anotador automático	Frequência de ocorrência como Nome Próprio
grilo	246	Nome	246
lazer	143	Nome	134
máxima	162	Adjectivo; Nome	87

Reunidos todos os dados atrás mencionados resultantes da desambiguação automática (INESC) e da desambiguação manual (CLUL) - 335.637 contextos analisados - procedeu-se à indexação final do léxico. (Cfr. infra "Amostragens da indexação final do Léxico")

Classificação morfossintáctica do Léxico

Os itens lexicais (lemas e formas) constituintes do léxico são acompanhados dos códigos de classificação que a seguir se apresentam.

¹⁸ *Grilo*, antropónimo, *Lazer* e *Máxima*, nomes de Revistas.

Nome_____	N	Conjunção_____	C
Verbo_____	V	Numeral_____	M
Adjectivo_____	A	Interjeição_____	I
Pronomes: pessoal _____	Pp	Estrangeirismo_____	Xf
demonstrativo _____	Pd	Abreviatura_____	Xa
indefinido _____	Pi	Acrónimo/Sigla _____	Xy
possessivo _____	Po	Símbolo_____	Xs
interrogativo _____	Pt	<i>Se</i> medio-passivo_____	U
relativo_____	Pr	Elemento de locução_____	L
exclamativo _____	Pe	Partículas enfáticas_____	E
reflexo _____	Pf	Grafia não-convencional _____	*
Artigo: definido _____	Td	Contracção_____	+
indefinido _____	Ti	Cabeça de lema _____	@
Advérbio_____	R	Cabeça de lema reconstituída por não ter ocorrido no <i>corpus</i> _____	[]
Preposição_____	S		

Estes códigos correspondem a categorizações em partes do discurso, com subcategorização para os Pronomes e Artigos, e a outras classificações como Estrangeirismo, Abreviatura e Acrónimo ou Sigla.

Considerou-se também como provável que algumas formas só tivessem ocorrido no *corpus* como elementos de locução. Assim foram analisadas em contexto e, nos casos pertinentes, anotadas com o código L (elemento de locução) formas como por exemplo: *acerca, apesar, aquando, cata, cima, conseguinte, desfavor, prol, redor, rés, riba, tocante, tona*, etc.

As formas com ortografias diferentes das que estão actualmente em vigor foram incluídas nos respectivos lemas, precedidas de asterisco.

Quadro com o número de lemas que ocorreram no *corpus* em cada uma das categorias consideradas:

Classificação	Nº de lemas diferentes	Frequência de ocorrência no CORLEX ¹⁹
N	14.515	4.115.080
V	4.154	2.417.516
A	6.284	1.060.376
Pp	23	128.115
Pd	19	276.556
Pi	38	309.714
Po	10	124.657
Pt	6	14.836
Pr	10	282.967
Pe	5	3.778
Pf	1	62.406
Td	2	2.235.633
Ti	2	299.172
R	993	2.218.976
S	29	2.946.700
C	32	895.426
M	57	159.537
I	71	12.108
Xf	533	43.538
Xa	24	9.250
Xy	134	51.302
Xs	4	1.037
U	1	4.681
L	30	296.286
E	3	4.117

Informação quantitativa

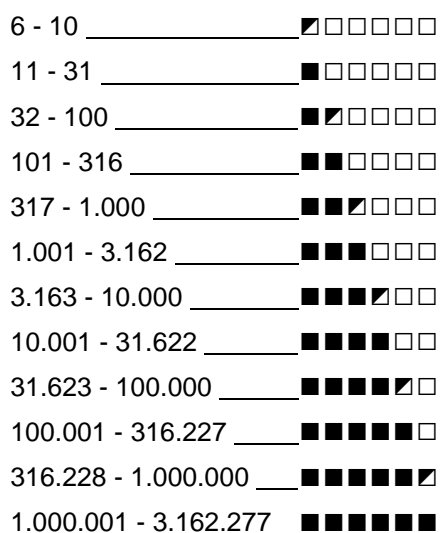
Como foi dito, os lemas que constam do Léxico atingiram frequência igual ou superior ao limiar 6.

¹⁹ Algumas formas podem ser contabilizadas mais do que uma vez, pois podem pertencer a mais do que uma entrada lexical (exemplo: *da* (Preposição + Artigo) com frequência 231.356 ocorre como forma do lema *de* e do lema *a*).

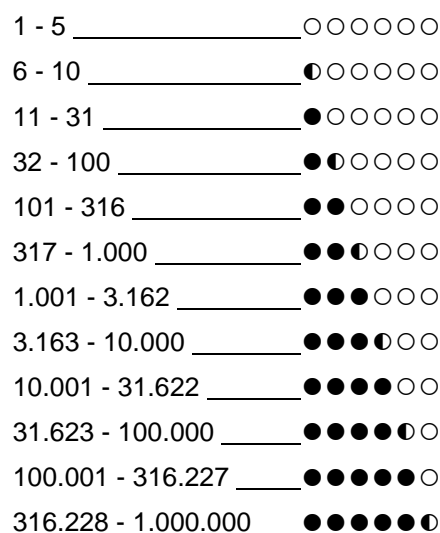
Junto de cada cabeça de lema e de cada forma dos lemas é apresentada uma informação de carácter quantitativo sobre a sua frequência no *corpus*. Uma vez que o intervalo de variação de ocorrência é muito grande, quer para as entradas, quer para as formas, utilizou-se uma escala logarítmica, a partir do logaritmo de base 10 ($\log_{10}/2$), para se obter uma distribuição mais uniforme dos dados quantitativos. Estes dados são representados por sequências de caracteres gráficos que indicam os seguintes valores:

Patamares de Frequência ($\log_{10}/2$):

Lemas:



Formas:



À coluna das Frequências segue-se uma coluna que informa sobre a repartição da ocorrência de cada forma em relação aos dois *sub-corpora*:

Repartição:

Ocorreu no *sub-corpora* oral _____ O
 Ocorreu no *sub-corpora* escrito _____ E
 Ocorreu no *sub-corpora* oral e no *sub-corpora* escrito _____ OE

A divulgação deste Léxico será feita muito em breve através da *internet* - na página do CLUL e na página do INESC e numa publicação em papel da Editorial VERBO.

Excerto do Léxico de Frequências

Por ordem alfabética:

@ MAÇÃ (N)	■■□□□□	@ MACADAME (XF)	■□□□□□
maçã (N)	●●○○○○ OE	macadame (Xf)	●○○○○○ E
maças (N)	●●○○○○ OE	macadames (Xf)	○○○○○○ E
maçãzinhas (N)	○○○○○○ E		
		@ MAÇADOR (A)	■□□□□□
@ MACABRO (A)	■■□□□□	maçador (A)	○○○○○○ OE
macabra (A)	●○○○○○ E	maçadora (A)	○○○○○○ OE
macabras (A)	●○○○○○ E	maçadoras (A)	○○○○○○ E
macabro (A)	●○○○○○ E	maçadores (A)	○○○○○○ OE
macabros (A)	○○○○○○ E		
		@ MAÇADOR (N)	■□□□□□
@ MACACO (A)	■□□□□□	maçador (N)	●○○○○○ OE
macaca (A)	○○○○○○ E	maçadora (N)	○○○○○○ OE
macacas (A)	○○○○○○ O	maçadoras (N)	○○○○○○ E
macaco (A)	●○○○○○ OE	maçadores (N)	●○○○○○ OE
macacos (A)	○○○○○○ OE		
		@ MACAENSE (A)	■□□□□□
@ MACACO (N)	■■□□□□	macaense (A)	●○○○○○ E
macaca (N)	●○○○○○ E	macaenses (A)	○○○○○○ E
macacas (N)	○○○○○○ O		
macaco (N)	●●○○○○ OE	@ MACAENSE (N)	■□□□□□
macacos (N)	●●○○○○ OE	macaense (N)	○○○○○○ E
macaquinha (N)	○○○○○○ E	macaenses (N)	●○○○○○ E
macaquinho (N)	○○○○○○ E		
macaquinhos (N)	○○○○○○ E	@ MACAMBÚZIO (A)	■□□□□□
macaquitos (N)	○○○○○○ E	macambúzia (A)	○○○○○○ E
		macambúzias (A)	○○○○○○ E
@ MAÇADA (N)	■■□□□□	macambúzio (A)	○○○○○○ E
maçada (N)	●●○○○○ OE	macambúzios (A)	○○○○○○ E
maçadas (N)	○○○○○○ OE		

@ MAÇANETA (N) ▣□□□□□
 maçaneta (N) ●○○○○○ E
 maçanetas (N) ○○○○○○ E

@ MAÇARICO (N) ■□□□□□
 maçarico (N) ●○○○○○ OE
 maçaricos (N) ○○○○○○ OE

@ MAÇÃO (N) ■□□□□□
 maçã (N) ●○○○○○ E
 mações (N) ○○○○○○ E

@ MAÇAROCA (N) ■□□□□□
 maçaroca (N) ●○○○○○ OE
 maçarocas (N) ●○○○○○ E

@ MACAQUICE (N) ▣□□□□□
 macaquice (N) ○○○○○○ E
 macaquices (N) ○○○○○○ OE

@ MACARRÃO (N) ■□□□□□
 macarrão (N) ●○○○○○ E

@ MAÇAR (V) ■□□□□□
 maça (V) ○○○○○○ E
 maçada (V) ○○○○○○ OE
 maçadas (V) ●○○○○○ OE
 maçado (V) ○○○○○○ E
 maçados (V) ○○○○○○ OE
 maçá-lo (V Pp) ○○○○○○ E
 maçam-me (V Pp) ○○○○○○ O
 maçar (V) ●○○○○○ OE
 maçaram (V) ○○○○○○ E
 maçaram-no (V Pp) ○○○○○○ E
 maçar-me (V Pp) ○○○○○○ E
 maçar-se (V Pf) ○○○○○○ OE
 maçar-te (V Pp) ○○○○○○ OE
 maças (V) ○○○○○○ E
 maçavam-no (V Pp) ○○○○○○ E
 mace (V) ○○○○○○ E
 macei-me (V Pp) ○○○○○○ E
 macem (V) ○○○○○○ E
 maço (V) ○○○○○○ OE
 maço-me (V Pp) ○○○○○○ E
 maçou (V) ○○○○○○ E

@ MACEDÓNICO (A) ■□□□□□
 macedónica (A) ○○○○○○ E
 macedónicas (A) ○○○○○○ E
 macedónico (A) ●○○○○○ E
 macedónicos (A) ○○○○○○ E

@ MACEDÓNIO (A) ■□□□□□
 macedónias (A) ○○○○○○ E
 macedónio (A) ●○○○○○ E
 macedónios (A) ○○○○○○ E

@ MACEDÓNIO (N) ■□□□□□
 macedónio (N) ○○○○○○ E
 macedónios (N) ●○○○○○ E

@ MACERAÇÃO (N) ■□□□□□
 maceração (N) ●○○○○○ E
 macerações (N) ○○○○○○ E

@ MACERAR (V) ■□□□□□
 macerada (V) ○○○○○○ E
 maceradas (V) ○○○○○○ E

macerado (V)	●○○○○○ E	@ MACHADADA (N)	■□□□□□
macerados (V)	○○○○○○ E	machadada (N)	●○○○○○ E
maceram (V)	○○○○○○ E	machadadas (N)	●○○○○○ E
macerando (V)	○○○○○○ E	machadadinhas (N)	○○○○○○ E
macerar (V)	●○○○○○ E		
macerassem (V)	○○○○○○ E		
macerava (V)	○○○○○○ E		
@ MACHADA (N)	■□□□□□		
machada (N)	○○○○○○ E		
machadinha (N)	○○○○○○ OE		
machadinhas (N)	○○○○○○ E		

Bibliografia

"Language Resources: PAROLE" (1998), in *Language Engineering Progress and Prospects' 98*, Luxembourg, Telematics Applications Programme, DGXIII-E-5, 118-119.

Bacelar do Nascimento, M.F., M.L. Garcia Marques e M.L. Segura da Cruz (1987), *Português Fundamental*, vol. II - *Métodos e Documentos*, tomo 1 - *Inquério de Frequência*, Lisboa, INIC, CLUL.

Bacelar do Nascimento, M.F., P. Rivenc e M.L. Segura da Cruz (1987), *Português Fundamental*, vol. II - *Métodos e Documentos*, tomo 2 - *Inquério de Disponibilidade*, Lisboa, INIC, CLUL.

Duncan, J. (1972), *Frequency Dictionary of Portuguese*, PhD Dissertation, Stanford, Stanford University.

Halliday, M.A.K. (1993), "Quantitative studies and probabilities in grammar", in Michael Hoey (ed.) *Data, Description, Discourse, Papers on the English Language in honour of John McH Sinclair*, London, Harper Collins Publishers, 1-25.

Português Fundamental, Vocabulário e Gramática, tomo 1 - Vocabulário (1984), Lisboa, INIC, CLUL.

Sinclair, J.M. (ed.) (1995), *Looking Up, An Account of the COBUILD Project in lexical computing*, London, Harper Collins Publishers.

Zampolli, A. (coord.) (1995), *Towards a Network of European Reference Corpora*, *Linguistica Computazionale*, vol. XI, Pisa, Giardini Editori e Stampatori.

Zipf, G.K. (1935), *The Psychology of Language*, Boston, Houghton Mifflin.