

## The use of concordancing in Portuguese teaching

Luísa Alice Santos Pereira  
 Centro de Linguística da Universidade de Lisboa (CLUL)  
<http://www.clul.ul.pt/>

### Introduction

The "*Corpus de Referência do Português Contemporâneo*", now with 153 Million words, is a *corpus* constituted by samples of Portuguese discourse, both spoken and written, representing all its national varieties, and those from Goa, Macao and East Timor. It is being developed as you can see in the following pages, in the **map** and in the **diagram**.

This *corpus* constitutes more and more the basis of many linguistic researches such as studies on different linguistic areas, theoretical and applied, lexicon construction and lexicographic work.

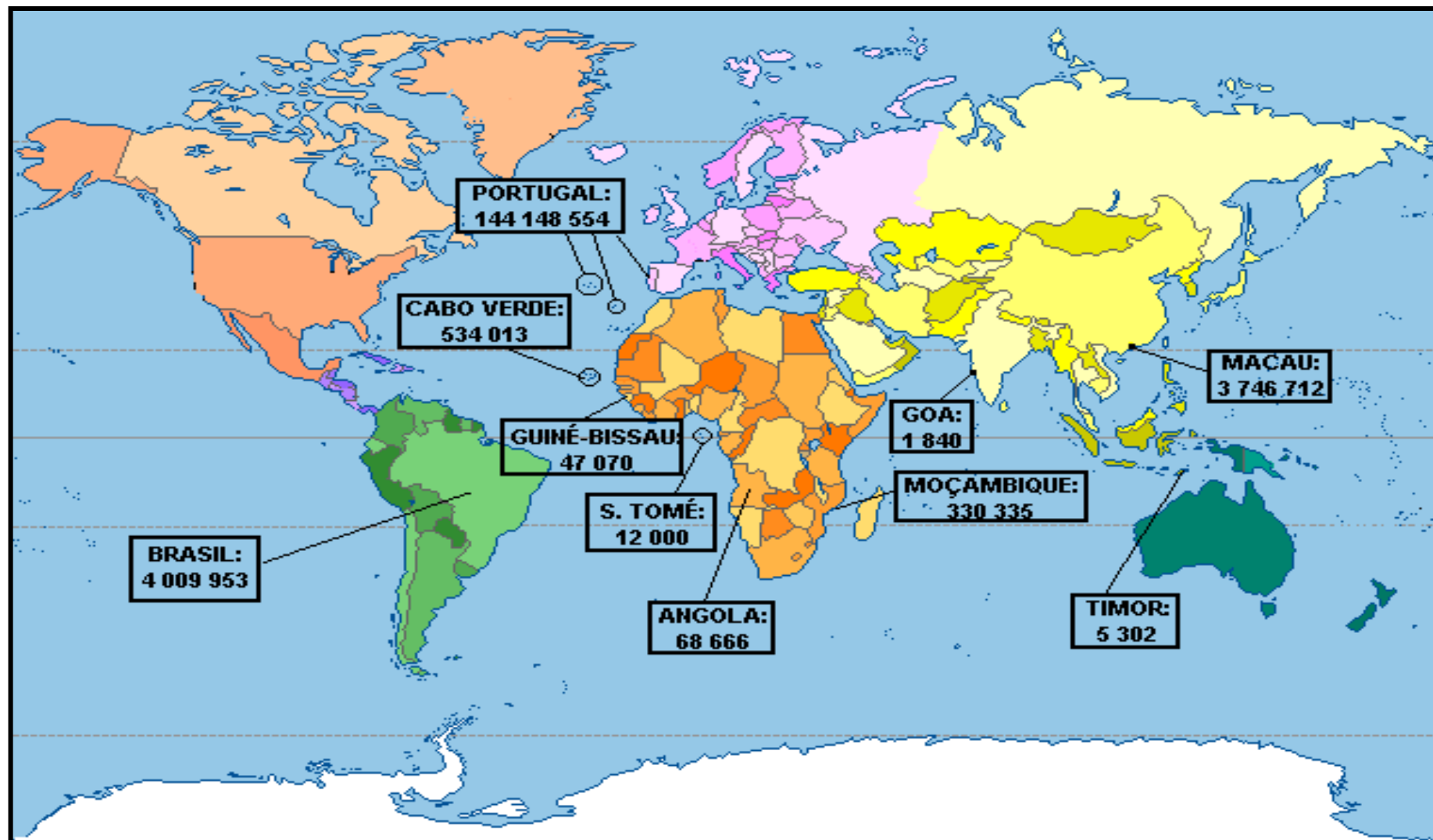
#### 1.

At the CLUL, the following materials have been developed, which are or will be soon available:

Available products, issued from CLUL or in partnership, based on the " <i>Corpus de Referência do Português Contemporâneo</i> " (CRPC)		
Published transcriptions (P.F.) of Portuguese spoken <i>Corpus</i>	CLUL	<a href="http://www.clul.ul.pt">http://www.clul.ul.pt</a>
" <b>Léxico Multifuncional Computorizado do Português Contemporâneo</b> "	partnership	<a href="http://www.clul.ul.pt">http://www.clul.ul.pt</a>
" <b>Português Falado</b> " - Records with transcription alignment	partnership	4 CD_ROM
3 million words of <i>corpus</i> PAROLE, from which 250000 morphosyntactically tagged with human desambiguation	partnership	ELRA catalogue - <a href="http://www.elda.fr">http://www.elda.fr</a>
PAROLE Lexicon with 20000 entries morphosyntactically tagged and syntactically described	partnership	ELRA catalogue - <a href="http://www.elda.fr">http://www.elda.fr</a>
3000 units from PAROLE Lexicon semantically characterised (SIMPLE programme) - multilingual	partnership	<a href="http://www.ub.es/gilcub/SIMPLE/simple.html">http://www.ub.es/gilcub/SIMPLE/simple.html</a>
Available in 2002		
" <i>Corpus</i> compartilhado VARPORT"	Partnership	
"Recursos linguísticos para o português"	Partnership	
<i>Corpus</i> REDIP	Partnership	
Available in 2003		
C-ORAL-ROM 4 romanic languages comparable spoken <i>Corpus</i>	Partnership	

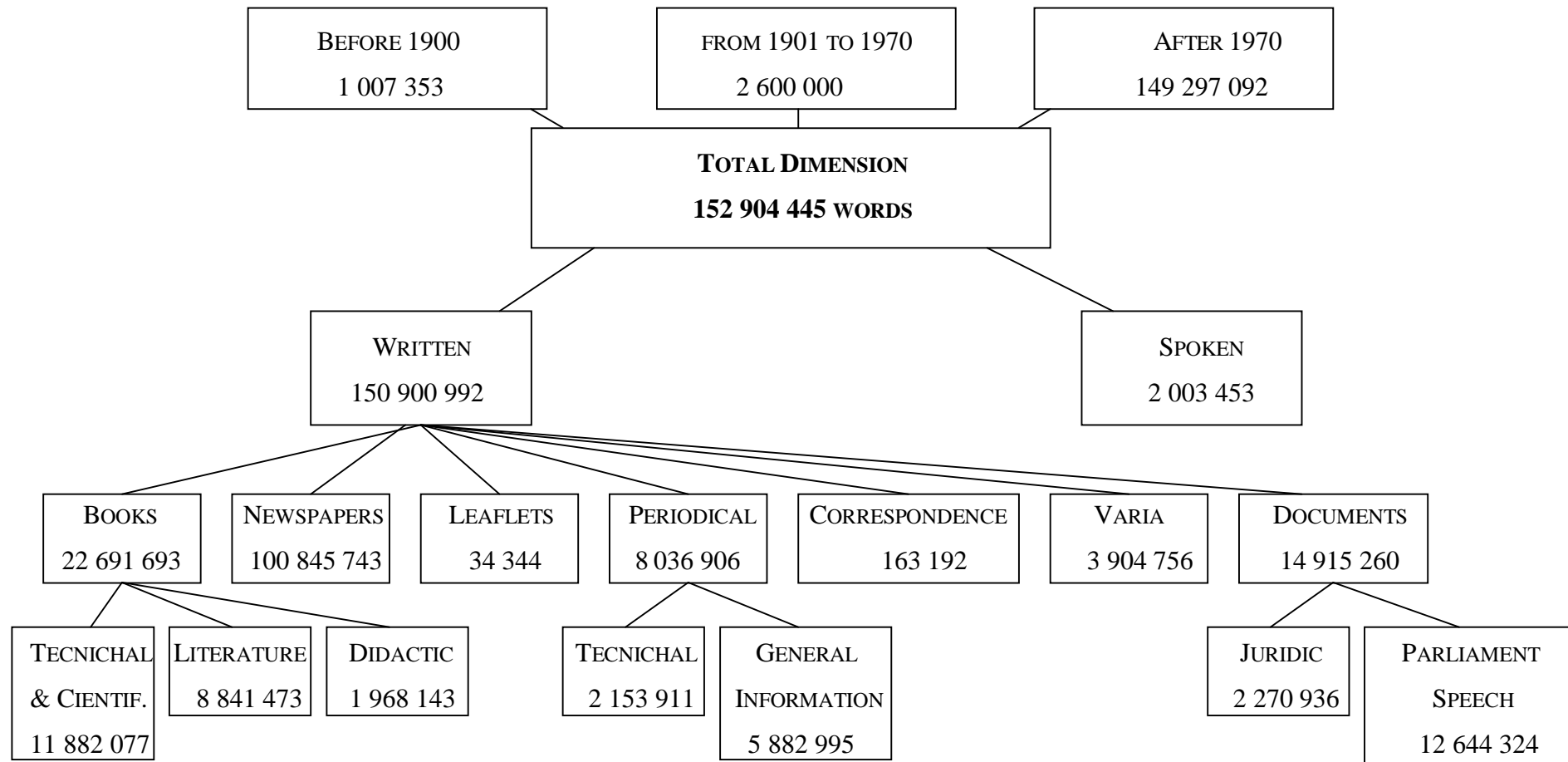
CORPUS DE REFERÊNCIA DO PORTUGUÊS CONTEMPORÂNEO - (CLUL)

DISTRIBUIÇÃO GEOGRÁFICA (Outubro de 2001)



Centro de Linguística da Universidade de Lisboa

***CORPUS DE REFERÊNCIA DO PORTUGUÊS CONTEMPORÂNEO***



From these materials, mainly focused are the "**Léxico Multifuncional Computorizado do Português Contemporâneo**" and "**Português Falado**".

The first is a general **Lexicon of Portuguese**, which contains 26980 lexical entries and respective forms (140976) with grammatical (morphosyntactical) and quantitative information. The **Lexicon** is constituted by the lexical entries that got a frequency of 6 or more, followed by all the respective forms (inflected forms and compounds). The level of *corpus* frequency (above 6) is given in a probabilistic basis.

This **Lexicon** is based on a 16.2 Million word contemporary Portuguese sub-*corpus*, written (15.35 M words) and spoken (0.86 M words), extracted from the CRPC. This sub-*corpus* was designed accordingly to the principles generally established and the international recommendations on the dimension and design of general use linguistic *corpora* for lexical extraction.

The "**Léxico Multifuncional Computorizado do Português Contemporâneo**" is a very useful product as a basis for different applications such as dictionaries, translation and NLP. It may also be an useful tool for teachers exploration. It is recently available on-line in PDF format by alphabetical and decrescent frequency order at the site of CLUL ([www.clul.ul.pt/](http://www.clul.ul.pt/)).

#### Alphabetical order of the lemmas

@ maca (N)	■□□□□	macabro (A)	●●○○○○
maca (N)	●○○○○○	macabros (A)	●○○○○○
macas (N)	●○○○○○		
@ maça (N)	■□□□□	@ macaco (A)	■□□□□
maça (N)	●○○○○○	macaco (A)	●○○○○○
maças (N)	●○○○○○	@ macaco (N)	■■□□□
@ maçã (N)	■■□□□	macaca (N)	●○○○○○
maçã (N)	●●○○○○	macacas (N)	○○○○○○
maças (N)	●●○○○○	macaco (N)	●●○○○○
maçãzinhas (N)	○○○○○○	macacos (N)	●●○○○○
@ macabro (A)	■■□□□	macaquinha (N)	○○○○○○
macabra (A)	●○○○○○	macaquinho (N)	●○○○○○
macabras (A)	●○○○○○	macaquinhos (N)	○○○○○○
		macaquitos (N)	○○○○○○

#### Decrescent frequency order of the lemmas

@ funerário (A)	■■□□□	-----	
funerários (A)	●●○○○○	russo-japonesa (N)	○○○○○○
funerárias (A)	●●○○○○	@ severo (A)	■■□□□
funerário (A)	●●○○○○	severo (A)	●●○○○○
funerária (A)	●●○○○○	severa (A)	●●○○○○
@ invisível (A)	■■□□□	severas (A)	●●○○○○
invisível (A)	●●○○○○	severos (A)	●●○○○○
invisíveis (A)	●●○○○○	severíssima (A)	○○○○○○
@ maçã (N)	■■□□□		
maçã (N)	●●○○○○	<b>LEGEND:</b>	
maças (N)	●●○○○○	<u>Frequency bands</u> ( $\log_{10}/2$ ):	
maçãzinhas (N)	○○○○○○	Lemmas:	
@ raciocínio(N)	■■□□□	6 - 10	■□□□□
raciocínio (N)	●●○○○○	11 - 31	■□□□□
raciocínios (N)	●●○○○○	32 - 100	■■□□□
@ russo (N)	■■□□□	101 - 316	■■□□□
russos (N)	●●○○○○	317 - 1000	■■■□□
russo (N)	●●○○○○	1001 - 3162	■■□□□
		3163 - 10000	■■□□□

10001 - 31622	■■■■■□□
31623 - 100000	■■■■■□□
100001 - 316227	■■■■■□□
316228 - 1000000	■■■■■□□
1000001 - 3162277	■■■■■□□

Forms:

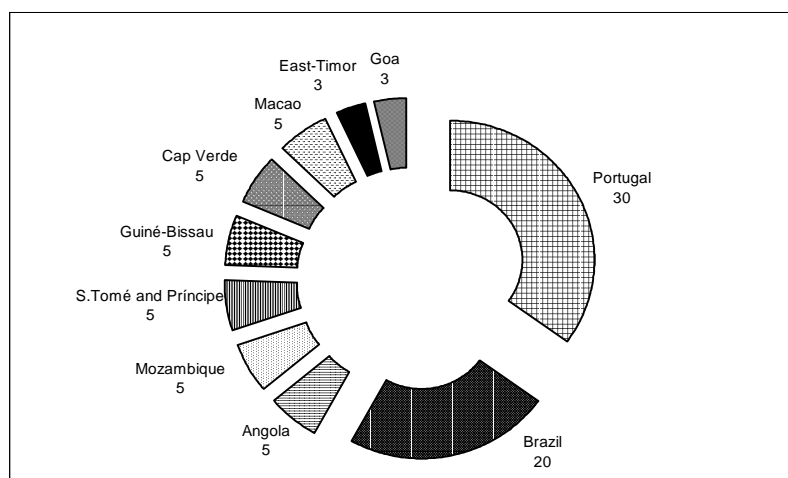
1 - 5	○○○○○
6 - 10	●○○○○
11 - 31	●○○○○
32 - 100	●●○○○
101 - 316	●●○○○
317 - 1000	●●○○○
1001 - 3162	●●○○○
3163 - 10000	●●○○○
10001 - 31622	●●○○○
31623 - 100000	●●○○○
100001 - 316227	●●○○○
316228 - 1000000	●●○○○

Codes:

Noun	N
Verb	V
Adjective	A
Pronoun	P
Article	T
Adverb	R
Preposition	S
Conjunction	C
Numeral	M
Interjection	I
Foreign word	F
Abreviation	X
Acronimous/Sigla	G
Symbol	B
Se medio-passive	U
Element of group	L
Emphasys Particle	E
Element out of order	_d
Non-conventional writing	*
Contraction	+
Head of lemma	@
Head of lemma reconstituted	
Because it didn't occur really	
in the corpus	[ ]

The "**Português Falado**" is the most recent edition of CLUL/Camões Institut. It is constituted by authentic spoken documents and it aims mainly to develop the capacity of Portuguese language understanding and production between foreign students of medium or high level Portuguese studies. The materials now published contribute to the observation and analysis of spoken Portuguese in its geographical varieties. Thus they are also very useful to teachers, translators, interpreters and researchers in general.

The 86 texts edited in CD\_ROM are text examples of spoken Portuguese varieties from Portugal (30), Brazil (20), Angola (5), Cape Vert (5), Guinea-Bissau (5), Mozambique (5), S. Tomé e Príncipe (5), Macao (5), East-Timor (3) and Goa (3).



These texts were recorded in situations of communication, from the more informal to more formal discourse, such as that of radio, for instance. The orthographic transcriptions and the alignment between sound and the corresponding graphical representation were performed from the recordings. Aiming at the user to be able to hear the recording and to read simultaneously the respective transcription in the computer screen, a coloured light runs over the transcription of the sequence which is being listened. The user can control what he is listening, can repeat sequences or jump parts of the text (Cf. GONÇALVES & VELOSO, 2000). It is published in 4 CD\_ROM available at CLUL (fbacelar.nascimento@clul.ul.pt) or at the Instituto Camões (ded@instituto-camoes.pt).

2.

Only these two products are focused because not only they are the most recent ones available from CLUL, but due to their usefulness to the *corpus*-based teaching. However, the *corpus* potential is endless and teachers can find in it a very significant source for new approaches to language teaching.

## 2.1.

As an example, and just giving frequency information, the comparative results of occurrences of the verb **deduzir**, observed in 6 CRPC sub-*corpora* of comparable dimensions (c. 1 M words) are presented. The analysis aims at observing the frequencies of one lemma and its forms, and also the occurrence of its syntactical, semantic and discursive uses. In this way, comparative tables of the verb **deduzir** in different discursive contexts, both spoken (SPK.) and written are shown: journalistic (JOURN.), literary (LIT.), didactic (DID.), economic (ECON.) and juridical (JUR.) (Cf. BACELAR do NASCIMENTO (under publ.c)).

Global frequencies of the verb **deduzir** in analysed *corpora*

CORPORA	SPK.	JOURN.	LIT.	DID.	ECON.	JUR.
Frequency of <b>deduzir</b>	6	9	7	57	40	242

Distribution of v. **deduzir** forms in analysed *corpora*

v. deduzir forms	Distribution per Corpus					
	Spk.	Journ.	Lit.	Did.	Econ.	Jur.
deduz		2		8		9
deduz-se	1	2	1	3		1
deduza						1
deduzem				2		1
deduzi		1	1			
deduzia			1			
deduziam	1					
deduzida		2		6	2	48
deduzidas				1	3	11
deduzido	1		1	1	6	56
deduzidos					8	
deduzimos				4		12
deduzindo				2	5	6
deduzir	1	2	1	21	13	35
deduzir-se					1	1
deduzirá						1
deduziram						9
deduzirem	1					
deduziria				1		
deduzirmos				2		
deduziste				1		
deduziu			1	5	1	50
deduziu-se						1
deduzo	1		1			
Totals	6	9	7	57	40	242

## 2.2.

Concordancing use can show syntactic / semantic productivity.

It can be seen that in some cases the syntactic places are fulfilled with repeated specific lexical items. An example for the verb **deduzir** is given, in which the juridical discourse assumes a special position. It is the repetitiveness of these words that contributes to show the importance of specialized *corpora*, e.g. when preparing courses to language students, mainly when these courses are designed for specific purposes.

Verb **DEDUZIR** observed on 6 *corpora* of  $\pm$  1M words each  
(spoken, journalistic, literary, didactical, economical, juridical)

(Syntactic and semantic information)

**Deduzir**  $\cong$  **To deduce**: syntactic structure and occurrence frequency in each *corpus*

Syntactic structure		SPK.	JOURN.	LIT.	DID.	ECON.	JUR.
V	N F (de, por, a partir de,... N) Que F	4	7	5	5	10	20

Ex:

<i>as carências são muitas e Nestas circunstâncias, é legítimo</i>	<i>deduz-se deduzir deduziu deduzidas deduzir deduz-se,</i>	<i>das suas palavras que também as potencialidades...(SPK.) que não se trata de um projecto consensual. (JOURN.) ele (LIT.) não são aplicáveis imediatamente (DID.) uma progressão mais moderada (ECON.) com suficiente clareza, que a indemnização devida...(JUR.)</i>
<i>Querem evitar a tropa, As conclusões aí a estabilização dos depósitos a prazo leva a e demais circunstâncias do caso concreto,</i>		

**Deduzir**  $\cong$  **To deduct**: syntactic structure and occurrence frequency in each *corpus*

Syntactic structure		SPK.	JOURN.	LIT.	DID.	ECON.	JUR.
V	N de, a, em N	2	0	1	7	29	23

Ex:

<i>chegar a casa com mais alguns dólares para Soma-se depois a despesa, Se a esta matéria orgânica O titular terá de somar os montantes que tinha o direito de</i>	<i>deduzirem deduz-se deduzirmos deduziu deduzir</i>	<i>daquela despesa (ORAL) do dinheiro entregue na véspera (LIT.) a que é gasta por esses operadores (DID.) no IRS (ECON.) na Retribuição do trabalhador o montante (JUR.)</i>
--	--	---

**Deduzir**  $\cong$  **To charge**: syntactic structure and occurrence frequency in each *corpus*

Syntactic structure		SPK.	JOURN.	LIT.	DID.	ECON.	JUR.
V	N (contra N)	0	2 <sup>1</sup>	0	0	0	222

<sup>1</sup> - In both occurrences, it is journalistic text on juridic themes

Ex:

<i>Na acusação O Ministério Público na comarca de Lisboa</i>	<i>deduzida deduziu</i>	<i>pelo Ministério Público, quatro dos arguidos... (JORN.) acusação contra os arguidos (JUR.)</i>
--	-------------------------	---

### 3.1.

The information above confirms that the high frequency of given lexical items can be determinant in the characterization of text types. As it is stated in BAAYEN (1998: 90) "Word frequencies constitute a rich source of information on the relation between lexis and text type. General summary measures capture global properties such as repetitiveness and vocabulary richness, but much more detailed information is available if we allow ourselves to consider individual words in specific frequency ranges". The specificity of texts and themes is reinforced by the observation of words which occur typically in homogeneous texts. The verb **deduzir** occurs as it follows in juridical discourse:



Another example which can offer good information is the realization of verb inflexion. If we observe in the Portuguese *corpus* the lemmas **arreigar-se**, **graduar** and **sofisticar** we can see that they are just used in the past participle in **spoken discourse** and that the **written discourse** has also a strong tendency to use them mainly in the past participle. Thus the *corpus* shows defectivities about which we have no information neither in grammars nor in dictionaries; both repeat, usually, the same traditional lists of defectives and they do not make reference to emergent uses (BACELAR do NASCIMENTO (2000a)).

VERB	Frequency		SPOKEN		WRITTEN	
	Spoken	written	Verbal tense	Form	Verbal tense	Form
<b>ARREIGAR-SE</b>	7	28	Past participle	Arreigado 3	Past participle	arreigada 11
				Arreigados 2		arreigado 6
				Arreigado 2		arreigados 6
						arreigadas 2
					Infinitive	arreigar-se 1
					Pret. Imperfeito	arreigava-se 1
<b>GRADUAR</b>	3	75	Past participle	Graduado 2	Past participle	graduado 28
				Graduados 1		graduados 21
						graduada 18
					Infinitive	graduar 3
					Simple past	graduou 2
<b>SOFISTICAR</b>	3	203	Past participle	sofisticada 1	Past participle	sofisticados 60
				sofisticado 1		sofisticado 56
				sofisticados 1		sofisticada 51
						sofisticadas 35
					Infinitive	sofisticar 1

Concordancies and collocations can also be very useful by shedding new light on the real uses of near-synonyms. The adjectives **célebre**, **famoso** and **notável** are presented in Portuguese dictionaries as synonyms. But, the concordancies and combinatories, extracted from a 12 M word *corpus* and organized according to the Mutual Information (MI), revealed very distinguished lexical patterns.

Examples are presented in three tables. In Table 1, the collocates are shown according to the **MI** decrescent order, considering the pair of words. Table 2 gives the collocates of **noun** category for each of the adjectives. Table 3 concerns the collocates of **adverb** category (Cf. BACELAR do NASCIMENTO (2000b)).

### Associative patterns of three "near-synonyms": Célebre - Famoso - Notável

**Table 1: MI decrescent order of the pair (considering the head of lemma)**

*** FT 454 CÉLEBRE ***			*** FT 686 FAMOSO ***			*** FT 433 NOTÁVEL ***		
Collocate	MI	Fq. of the pair	Collocate	MI	Freq. of the pair	Collocate	MI	Freq. of the pair
TRISTEMENTE	8.825	10	TORNAR	7.043	14	CONJUNTO	6.641	6
CRIMINOSO	8.320	4	NOME	6.876	11	QUALIDADE	6.213	6
FICAR	6.203	14	COLECÇÃO	6.429	4	VERDADEIRAMENTE	6.184	6
FRASE	5.891	7	FICAR	5.311	9	ESFORÇO	5.575	8
TORNAR	5.846	12	AMERICANO	5.253	4	OBRA	5.044	7
AUTOR	5.034	6	GENTE	4.262	5	ÉPOCA	4.840	6
MAIS	3.695	37	SUA	4.259	18	EXEMPLO	3.889	5
TÃO	3.561	10	DE	4.068	433	MAIS	3.849	48
SER	3.394	45	SER	3.953	67	SER	3.836	68
DE	3.259	325	MAIS	3.720	77	FAZER	3.792	7
SUA	3.186	11	MENOS	3.630	4	MUITO	3.769	16
EM	3.025	73	GRUPO	3.513	4	TRABALHO	3.622	7
SEUS	2.973	6	POR	3.501	55	TER	3.580	8
POR	2.901	31	TÃO	3.467	8	DE	3.393	254
MUITO	2.531	10	SEU	3.395	22	COM	2.480	29
DIA	2.449	4	CASA	3.236	4	E	2.388	77
JÁ	2.046	6	JÁ	3.087	23	EM	2.092	56
QUE	1.847	42	EM	3.079	64	PARA	1.902	25
E	1.816	46	MUITO	2.660	9			
COM	1.639	14	COMO	2.489	16			
			A	2.450	253			
			E	2.441	119			
			COM	2.122	30			
			NÃO	1.990	36			

**Table 2: Collocates - NOUNS**

CÉLEBRE	FAMOSO	NOTÁVEL
	<b>NOUNS</b>	
CRIMINOSO	NOME	CONJUNTO
FRASE	COLECÇÃO	QUALIDADE
AUTOR	AMERICANO	ESFORÇO
DIA	GENTE	OBRA
	GRUPO	ÉPOCA
	CASA	EXEMPLO
		TRABALHO

**Table 3: Collocates - ADVERBS**

CÉLEBRE	FAMOSO	NOTÁVEL
	ADVERBS	
TRISTEMENTE	MAIS	VERDADEIRAMENTE
MUITO	MENOS	MAIS
MAIS	TÃO	MUITO
TÃO	JÁ	
JÁ	MUITO	
	NÃO	

It can be seen that the lexical co-occurrences are quite different. On the one hand, each adjective shows to be used with relatively specific nouns and adverbs. On the other hand, the information provided shows that those nouns and adverbs are not inter-exchangeable.

When considering the collocate nouns it can be seen that it is used **frase célebre** but not "frase notável", **notável qualidade** but not "famosa qualidade" and **nome famoso** but not "nome notável". Relatively to the collocate adverbs, **célebre** collocates with tristemente, which has a semantical negative weight, and **notável** collocates with the more emphatic verdadeiramente. The other adverbs are normally used in the formation of comparative and superlative adjectives.

## Conclusion

To offer teachers and students the possibility to explore the language that they are studying, either as mother language, or second, or foreign language, it may be very useful and motivating.

The use of concordancing, collocations in particular, provides to the teachers very good sources to study and design materials for classroom work, either for preparing presentations or to organise exercises. Similarly, for students it constitutes a better way of understanding the language they are studying on different aspects, such as meaning and semantic disambiguation, real differences of near-synonyms, morphosyntactic classification, and information on the real use of terms in specialized discourses.

## Bibliography

- BAAYEN, R.H. (1998), "Lexis, Word Frequencies and Text Types", *IV-V Jornades de Corpus Lingüistics 1996-1997*, Barcelona, Institut Universitari de Lingüística Aplicada: Universitat Pompeu Fabra, pgs. 87-102.
- BACELAR do NASCIMENTO, Maria Fernanda (2000a) "*Corpus de référence du portugais contemporain*", in BILGER, Mireille (ed.) *Corpus. Méthodologie et applications linguistiques*, Paris, Honoré Champion, pp. 25-29.
- BACELAR do NASCIMENTO, Maria Fernanda (2000b) "Exemples de combinaisons établies pour l'écrit et pour l'oral à Lisbonne", in BILGER, Mireille (ed.) *Corpus. Méthodologie et applications linguistiques*, Paris, Honoré Champion, pp. 237-261.
- BACELAR do NASCIMENTO, Maria Fernanda (under publ.a) "O papel dos *corpora* especializados na criação de bases terminológicas", in CASTRO, Ivo & DUARTE, Inês (eds.) *Razões e Emoção. Miscelânea de estudos oferecida a Maria Helena Mira Mateus pela sua jubilação*.
- BACELAR do NASCIMENTO, Maria Fernanda (under publ.b) "Para um banco de dados do português falado e escrito. *Corpora* linguísticos: desenvolvimentos e aplicações", 1<sup>st</sup> International Meeting of AILP - International Association of Portuguese Linguistics, Lisbon, FLUL, October 8<sup>th</sup>, 2001.
- BACELAR do NASCIMENTO, Maria Fernanda (under publ.c) "Fenómenos de Lexicalização no Português Contemporâneo", AATSP, San Francisco, U.S.A., July 5th-9th, 2001.
- GONÇALVES, José Bettencourt & VELOSO, Rita (2000) "Spoken Portuguese: Geographic and Social Varieties", in GAVRILIDOU, M., CARAYANNIS, G., MARKANTONATOU, S., PIPERIDIS, S. & STAINHAOUER, G. (eds.) *LREC2000 Second International Conference on Language Resources Evaluation*, Proceedings, Volume II, ELRA, Athens, Greece.
- SINCLAIR, J., (1991) *Corpus, Concordances, Collocations*, Oxford, OUP.