An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications

Luísa Alice Santos Pereira, Amália Mendes
Centro de Linguística da Universidade de Lisboa (CLUL)
Complexo Interdisciplinar, Av. Prof. Gama Pinto, 2
1649-003 Lisbon Portugal
luisa.alice.sp@clul.ul.pt, amalia.mendes@clul.ul.pt

Abstract
This paper discusses a lexicographic approach to collocations, presenting the methodology, options, results and applications of an electronic Dictionary of Portuguese Collocations (DCP). The methodology underlying the dictionary involves the extraction from a corpus of contemporary Portuguese of lexical associations of pairs of word forms, contiguous or not. The significance of the pairs is statistically measured by the Mutual Information (MI) calculus, as well as by the MI weighted by the frequency of the pair (MIF). Other issues are discussed: frequency of the word forms vs. frequency of the lemmas, the organization of the collocations in the dictionary, grammatical patterns as source of lexical information, as well as the splitting of collocations into sense-groups.

1 Introduction

The electronic Dictionary of Portuguese Collocations (DCP) is a project being developed at the Center of Linguistics of the University of Lisbon (CLUL) [Pereira 1994; Bacelar do Nascimento 1998]. The goal of this dictionary is the constitution of an inventory of the most frequent lexical associations of contemporary European Portuguese, extracted from a 12M token corpus, which is a balanced sub-corpus of the *Reference Corpus of Contemporary Portuguese* of 150M [Bacelar do Nascimento 2000].

The concept of collocation is defined by Firth [1955] as the characterization of a word according to the words that typically co-occur with it. Firth's work [Firth 1957] awoke interest on the study of lexical co-occurrences and showed that the meaning of a word is closely related to the set of co-occurring words. It is becoming obvious that natural languages follow more regular patterns at syntagmatic level than they were believed to, and the study of corpus data allows us to identify those patterns. The identification of such associative patterns of the word gives important information on the meanings of the word and its actual uses [Sinclair 1991].

These associative patterns are considered an extension of Firth's notion of collocation. Thus this concept is extended by considering collocations to be associative models, showing a varying degree of fixedness, which excludes free association and ends in idiomatic chunks of words. Our purpose is to identify associative patterns that allow defining the word:

- by its relationship with systematically co-occurring lexical units;
- by its relationship with morphosyntactic and syntactic features: certain words always co-occur with a certain verb class, with specific temporal verb forms or with certain constructions;

- by extra-linguistic relationships (situational, contextual) related to different types of discourse (strong associations in one language register can be a weak association in another register) [Bacelar do Nascimento 1998].

The methodology followed to build the dictionary, mainly, the extraction and ordering of collocations is discussed in section 2. In section 3, are analysed some specific issues concerning the results obtained and further improvements to the project, such as the difference between lexical and grammatical patterns and the implementation of an automatic process of selection of the significant collocations. The final section reviews the importance of the associative patterns for a better understanding of the lexical, semantic, syntactic and pragmatic properties of lexical items and the applications of such results.

2 Extracting collocations from the corpus

The first step was to extract from the corpus all the pairs of words as well as all the groups of 2, 3, 4 and 5 words with frequency ≥ 2, using CLUL's software. The study of these larger groups proved that it was preferable to work with pairs of words. One of the problems encountered was the possible lexical variation of one or more members of the group, producing a large amount of multi-word units, most of them non-significant ones.

It was thus decided to work only with pairs of words, either contiguous or separated by 2 or 3 words. The frequency of each pair in the corpus was calculated and a statistical measure of the significance of the pair called Mutual Information (MI) was implemented. The MI is based on the frequency of the pair of words in the corpus and crosses this frequency with the isolated frequency of each word of the pair in the corpus [Church & Hanks 1990]. A sample of the pairs of the lemma *notável* 'remarkable' is presented in Table 1. In the first row, the lemma is preceded by information concerning its total frequency (FT) in the corpus. The table lists some of the pairs formed by the lemma (either the singular form *notável* or the plural form *notáveis*) and its collocates, ordered by the MI. Mutual Information is applied to the word forms (and not to the lemma) since the MI can be particularly high with one word form and not with another. For example, the lemma *pressão* 'pressure' is strongly associated to a small list of adjectives but only when occurring in the plural form *pressões* 'pressures' (*altas pressões* 'high pressures' is a strong association in weather reports).

However, the high number of word forms existing in Portuguese, especially in the case of verbs, led us to organize the word forms of the collocates in the dictionary under their lemma. The next step was, thus, to lemmatize all the word forms of the list of pairs of words. Consequently, the pairs (with their MI) were then reordered under their lemmas. For example, for the lemma *notável,* the dictionary provides information on several collocates, one of them being the lemma *conjunto* 'set, group', in both singular and plural word forms, as shown in Table 2. The first line in Table 2 presents the total frequency (FT) of the lemma *notável.* The second line identifies the collocate *conjunto* (lemma) and its frequency (frequency 6) in the corpus when occurring with *notável.* Under the lemma *conjunto* are then grouped the pairs of the word forms of both words and the MI of the pair (the pair *conjuntos notáveis* 'remarkable sets' has an MI of 6.641). For each pair of word forms, the dictionary presents the contexts of the corpus in KWIC format.

## *** FT 433 NOTÁVEL *** ##			
PAIR	MI	PAIR	MI
conjuntos notáveis	6.641	das notáveis	2.851
notáveis qualidades	6.213	mais notável	2.577
verdadeiramente notável	6.184	é notável	2.562
notável esforço	5.575	com notável	2.480
obras notáveis	5.044	tem notável	2.426
época notável	4.840	notáveis e	2.388
notável conjunto	4.255	notáveis do	2.337
obra notável	3.975	notável de	2.289
notável exemplo	3.889	notável e	2.176
notáveis mais	3.849	notáveis de	2.174
são notáveis	3.836	fazer notável	2.112
fez notável	3.792	notáveis no	2.092
muitos notáveis	3.769	no notável	2.085
feito notável	3.749	notáveis da	1.938
notável qualidade	3.646	na notável	1.936
notável trabalho	3.622	notáveis para	1.902
têm notável	3.580	do notável	1.886
notável foi	3.560	notável para	1.834
notáveis dos	3.393	da notável	1.795
foram notáveis	3.105	com notáveis	1.753
muito notável	3.030	é notáveis	1.342

Table 1: Sample of the pairs containing the word *notável* 'remarkable' ordered by the MI

```
## *** FT 433 NOTÁVEL *** ##
## ***  6 CONJUNTO (real:6) *** ##
# conjuntos notáveis # 6.641
#
# 2 conjuntos notáveis 1
   46485686 ta da Bacalhoa, destacam-se dois      conjuntos notáveis:      o da Casa do
  46485679  XV, conservando-se ainda muitos       conjuntos notáveis,      onde as duas

# notável conjunto # 4.255
#
# 2 notável conjunto 1
  159726773 i rentabilizar a presença de tão       notável conjunto        de guitarristas
   159726780 ciar a actividade política de um      notável conjunto        de intelectuais
#
# 2 conjunto notável 2
  46468365 ntam as decorações de lavores. O     conjunto mais notável     deste padr
  46468358 ejos sublinham a arquitectura. O     conjunto mais notável     e espectac
```

Table 2: Collocates of the lemma *notável* after lemmatization

The electronic format of the dictionary allows us to provide more information on the real uses of the collocations since there is no limitation of space. The dictionary presents all the contexts in which each collocation occurs in the corpus, and the dimension of these contexts can be larger or smaller according to the needs of the user. The lemmas of the collocates are then ordered according to the higher MI encountered. Finally, the lexical associations extracted and ordered are manually revised and the non-significant pairs of word forms are eliminated; the concordances in KWIC format are also manually revised and contexts that do not refer to the pair in question (sometimes due to punctuation) are eliminated.

This last step of the process is still under development, since we are aiming to rely more and more on automatic statistical processes for the elimination of the non-significant lexical associations. However, some of the word forms put forward by the MI calculation are not the most interesting collocates. Some of the most significant associations, according to the MI, are in fact the first and last name of personalities with high frequency in the corpus. To avoid this result, it is possible to add another calculus, the MIF, where the Mutual Information is weighted by the frequency of the pair [Baugh & Jellis 1996]. Thus, the pairs with both a high MI and a high frequency in the corpus will be identified as the most significant ones. Table 3 presents the most significant collocates of the lemma *pressão* 'pressure' (in both singular and plural form, respectively *pressão* and *pressões*) according to the MI and the MIF.

	MI
pressões pró-amnistia	10.341
pressões inflacionistas	9.424
pressão subglótica	9.302
limitadora pressão	9.302
pressão conjugavam-se	9.302
pressão origando	9.302
demissionários pressões	8.954
pressões equatoriais	8.874
pressões subtropicais	8.731
pressão 1015	8.608

	MIF
altas pressões	381.069
pressão atmosférica	341.912
baixas pressões	263.219
pressão arterial	214.645
pressão exercida	198.902
as pressões	183.958
forças pressão	158.713
aumento pressão	105.069
pressões subtropicais	78.581
grupos pressão	74.631

Table 3: Sample of the most significant collocates of the lemma *pressão* 'pressure' according to the MI and the MIF

In fact, when looking at Table 3, one can see that only one collocate is identified by both statistical measures: the word form *subtropicais* 'subtropical'. The fact that the MI considers collocates with very low isolated frequency in the corpus to be the more significant ones can be misleading, like in the case of the collocation *pressão origando* in line 6 of Table 3, where the collocate *origando*, a typo, is selected as highly significant. Collocations selected by the MIF are considered by native speakers of Portuguese as more conformant to their intuitions on which collocates of the lemma *pressão* are significant ones. However, the MIF calculus raises the exact opposite problem by giving more value to collocates with high isolated frequency in the corpus. One consequence is that the higher values of the MIF are mostly attributed to collocations with grammatical words. Although the case of the lemma *pressão* is not a good example of this, one of the higher values of MIF is given to the collocation *as pressões* 'the pressures' in line 6, where the collocate of *pressões* is an article.

3 What is a significant collocation?

As referred in the previous section, it is important to implement processes that allow us to filter the inevitable noise that is found in a list of the word pairs ≥ 2 in the corpus. Although the MI and the MIF provide an ordering of this list, the next step is to establish a cut-off point that will separate the significant collocations from the non-significant ones without loosing important information. The elimination of pairs of words separated by punctuation is a possible measure to reduce the need for manual intervention at the last step of our work. However, it is possible for significant collocations to be separated by an adverbial element. For example, the collocation *conjunto notável* 'remarkable set/group' (See Table 2) could occur in the following context: *conjuntos, sem dúvida, notáveis* 'sets/groups, with no doubt, remarkable'. A final decision in this matter will have to weight the loss of information and the gain in automatic process.

Another possible way to eliminate non-significant collocations would be to select only the pairs that are particularly frequent in one specific position. However, the results of this process are questionable if we look at the lemma *notável* and its collocate *conjunto* in Table 2, above. This pair occurs in different positions with the same frequency; nevertheless, this is a significant collocation in Portuguese, which should be selected.

Another issue regarding the significance of collocations is the difficulty in distinguishing between lexical and grammatical information. Although our initial aim is to achieve a list of the most frequent lexical associations of European Portuguese, we kept grammatical co-occurrents separated by one or more words, and thus showing interesting lexical information inside the window. Data in Table 4 show a fixed grammatical sequence beginning with the word *por* 'by' and finishing with the word *adiante* 'ahead' with one position filled by different lexical items.

ADIANTE 'ahead'

64 POR 'by'

pela adiante 'by ahead' 3.231 (MI)

	17 pela adiante 2	
via lá dentro, agitação, falácia	pela **casa** adiante.	Era cedo ainda
doutor@i não se fazia rogado; e,	pela **história** adiante	ia metendo
-se dele e do Oliveira, e correu	pela **igreja** adiante,	em direcção
her? O ódio vai diluindo e passa	pela **noite** adiante...	A mãe, da
e aí permanecem o mais do tempo,	pela **noite** adiante,	até o baile
seus braços! Interminàvelmente	pela **noite** adiante,	dormindo ou
alojar Deus na sua própria casa.	pela **rua** adiante,	aqui e ali, na
Depois demos o braço e fomos	pela **rua** adiante	calados e unido
cujo nome me fugia, a rebolar-se	pela **rua** adiante),	eram apenas a
escrita da navalha) prolonga-se	pela **tarde** adiante,	fala de pequ
a cair cada manhã, demoram-se	pela **tarde** adiante,	só com a noi
três noites metidos num comboio.	pela **viagem** adiante	já havia san

Table 4: Semantic patterns revealed by the inner position of the window of a pair

The pair *por adiante* 'by ahead' separated by one position shows semantic patterns concerning the type of lexical units filling this position, namely names referring either to time (*noite* 'night', *tarde* 'afternoon', *história* ‘story’) or space (*casa* 'house', *igreja* 'church', *rua* 'street') or both (*viagem* 'travel').

Another important syntactic information revealed by the associative patterns concerns verbal, nominal and adjectival subcategorization. Associative patterns formed by a lexical word and a grammatical word, like *hipótese de* ‘hypothesis of’ and *consiste em* ‘consists of’, can prove to be extremely useful for teaching Portuguese language.

The present results included in the dictionary have suffered few restrictions of information, considering the important syntactic and semantic information provided to the users by the contexts. The dictionary is the result of an automatic process of extraction of collocations, lemmatization and ordering. However, one of the objectives of this project is to make explicit the syntactic, semantic and pragmatic information that are now implicit through the contexts presented. Several developments of the project are now under consideration: automatic selection of significant collocates, disambiguation of different word classes, splitting collocations into different sense-groups and dissemination of results.

4 Collocations as a source of semantic information

The discussion of Table 4 showed the importance of collocations for uncovering semantic patterns. In fact, the different possible collocates of a lemma provide crucial information on the semantic properties of the lemma, mainly information on aspects like close synonyms, polysemy and terminology.

One of the interesting information provided by this type of dictionary concerns close synonyms that are presented in usual dictionaries as absolute synonyms, although the study of their lexical associations show important differences in the way speakers use these words. The word *notável* 'remarkable', already discussed, receives several synonyms in the Portuguese dictionaries, like *célebre* 'famous' and *famoso* 'famous', but show different collocational patterns reflecting semantic variations:

FT 454 CÉLEBRE 'famous'	**FT 686 FAMOSO 'famous'**	**FT 433 NOTÁVEL 'remarkable'**
	CO-OCCURENTS:	
CRIMINOSO ‘criminal’ (freq: 4)	NOME ‘name’ (freq: 11) .	CONJUNTO ‘group/set’ (freq: 6)
FRASE ‘sentence’ (freq: 7)	COLECÇÃO ‘collection’(freq: 4)	6 QUALIDADE ‘quality’ (freq: 6)
AUTOR ‘author’ (freq: 8)	AMERICANO ‘american’ (freq: 4)	8 ESFORÇO ‘effort’ (freq: 8)
DIA ‘day’ (freq: 4)	GENTE ‘people’ (freq: 5)	7 OBRA ‘work/production’(freq: 7)
	GRUPO ‘group’ (freq: 4)	6 ÉPOCA ‘times’ (freq: 6)
	CASA ‘house’ (freq: 4)	5 EXEMPLO ‘example’ (freq: 5)
		7 TRABALHO ‘work’ (freq: 7)

Table 5: Collocates of three words treated as synonyms in Portuguese dictionaries

The highlighting of the polysemy of lexical units is also one of the most productive application of collocation patterns, since the collocates of a word point towards different meanings of that word. In Tables 6, 7, 8 and 9, the lemma *pressão* 'pressure' has been split into different sense-groups. Table 6 regards weather reports contexts and points towards two collocates, the adjectives *subtropicais* 'subtropical' and *atmosférica* 'atmospheric', while Table 7 identifies the medical use of the word *pressão*, in the collocation *pressão arterial* 'arterial pressure'.

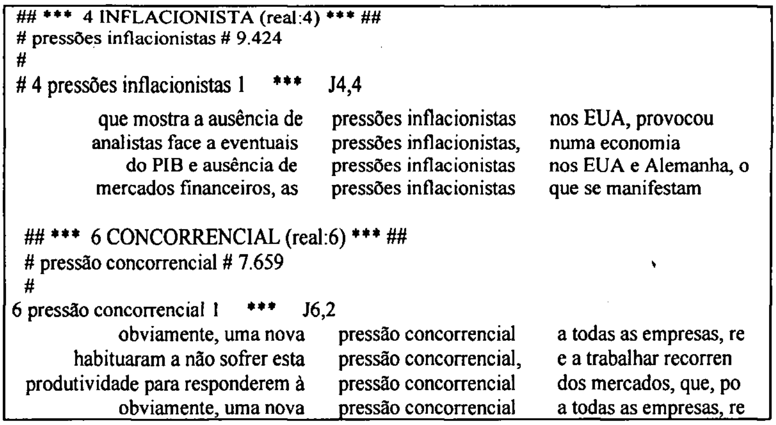
```
## *** 9 SUBTROPICAL (real:9) *** ##
# pressões subtropicais # 8.731
#
# 9 pressões subtropicais 1   ***   E9,2
          que provêm das altas   pressões subtropicais.     Na região equatorial,
       vidos por centros de altas   pressões subtropicais      sendo o anticiclone dos
      à direita, origina as altas   pressões subtropicais      pela subsidência. As
      pela subsidência. As altas   pressões subtropicais,     funcionando como
       ar polar alimenta as altas   pressões subtropicais      juntamente com o
#
## *** 44 ATMOSFÉRICO (real:46) *** ##
# atmosférica pressão # 8.643
#
# 40 pressão atmosférica 1   ***   T1,1  E37,6  U1,1  J1,1
       barómetros, para medir a   pressão atmosférica;    manómetros, para quanti
    Havia qualquer mudança de   pressão atmosférica.    De resto, a Califórnia or
       provar a sua existência?   pressão atmosférica     Os resultados obtidos nas
       xercida pelo ar chama-se   pressão atmosférica.    Pois claro! Reparaste qu
       permitem concluir que a   pressão atmosférica     se exerce em todos os se
```

Table 6: The lemma *pressão* 'pressure' in weather context

```
### *** 26 ARTERIAL (real:26) *** ##
# pressão arterial # 8.256
#
# 26 pressão arterial 1   ***   R20,7  T3,2  E2,2  J1,1
     sanguíneos e faz baixar a   pressão arterial.   influência da radiação solar s
      aorta, que tem o nome de   pressão arterial    e que se pode determinar por
       que contactam. 56. 1 - A   pressão arterial    resulta das forças de pressão
   Dos pesos; a determinação da   pressão arterial    (valores de 18 mm, para a má
  e severa, palpitações, queda da   pressão arterial    e náuseas. Nitrocelulose. É
```

Table 7: The lemma *pressão* 'pressure' in medical context

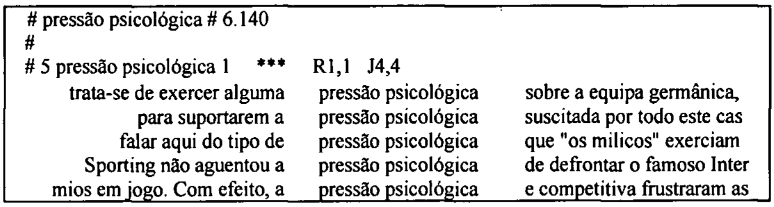
Two other meanings of *pressão* are uncovered in Tables 8 and 9. Table 8 refers to economical context with two collocations *pressão inflacionista* 'inflation pressure' and *pressão concorrencial* 'competitive pressure' as multi-words economical terms. Table 9 shows extension of the meaning of pressão to the domain of emotions with the collocation *pressão psicológica* 'psychological pressure'.



```
## *** 4 INFLACIONISTA (real:4) *** ##
# pressões inflacionistas # 9.424
#
# 4 pressões inflacionistas 1  ***   J4,4
      que mostra a ausência de   pressões inflacionistas    nos EUA, provocou
      analistas face a eventuais   pressões inflacionistas,   numa economia
            do PIB e ausência de   pressões inflacionistas    nos EUA e Alemanha, o
      mercados financeiros, as   pressões inflacionistas    que se manifestam

 ## *** 6 CONCORRENCIAL (real:6) *** ##
 # pressão concorrencial # 7.659
 #
6 pressão concorrencial 1  ***  J6,2
              obviamente, uma nova   pressão concorrencial    a todas as empresas, re
       habituaram a não sofrer esta   pressão concorrencial,   e a trabalhar recorren
  produtividade para responderem à   pressão concorrencial    dos mercados, que, po
              obviamente, uma nova   pressão concorrencial    a todas as empresas, re
```

Table 8: The lemma *pressão* 'pressure' in economical context



```
 # pressão psicológica # 6.140
 #
 # 5 pressão psicológica 1  ***   R1,1  J4,4
        trata-se de exercer alguma   pressão psicológica   sobre a equipa germânica,
               para suportarem a   pressão psicológica   suscitada por todo este cas
             falar aqui do tipo de   pressão psicológica   que "os milicos" exerciam
        Sporting não aguentou a   pressão psicológica   de defrontar o famoso Inter
       mios em jogo. Com efeito, a   pressão psicológica   e competitiva frustraram as
```

Table 9: The psychological meaning of the lemma *pressão*

Conclusions

The electronic Dictionary of Portuguese Collocations provides an inventory of the most significant lexical collocations in Portuguese. Collocations are selected based on frequency and statistical information. For each collocation, the DCP provides the full amount of contexts in which the collocations occur in the corpus. These real contexts allow the users to induce the syntactic, semantic and contextual properties of the multi-word forms.

The users of the dictionary are provided with information on observable preferences of lexical associations treated statistically, which will help them to distinguish significant factors from pure noise and to isolate phenomenon and induce generalizations.

The further developments of the DCP include the improvement of automatic processes identifying significant collocations, as well as providing explicit information on the properties that can, for now, be induced from the contexts.

References

[Bacelar do Nascimento 1998] Bacelar do Nascimento, M. F., 1998. *Dicionário de Combinatórias do Português,* Final Report. Centro de Linguística da Universidade de Lisboa, Lisbon.

[Bacelar do Nascimento 2000] Bacelar do Nascimento, M. F., 2000. 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, in: E. Gärtner et al. (eds.) *Estudos de Gramática Portuguesa (I),* pp. 185-200. Biblioteca Luso-Brasileira, Centro do Livro e do Disco de Língua Portuguesa, Frankfurt am Main.

[Baugh et al. 1996] Baugh, S., A. Harley & S. Jellis, 1996. The Role of Corpora in Compiling the Cambridge International Dictionary of English, in: *International Journal of Corpus Linguistics*, Vol. 1 (1), pp. 39-47. John Benjamins, Amsterdam.

[Church & Hanks 1990] Church, K. W. & P. Hanks, 1990. Word association norms, mutual information, and lexicography, in: *Computational Linguistics*, 16 (1), pp. 22-29.

[Firth 1955] Firth, J., 1955. Modes of meaning, in: *Papers in Linguistics 1934-1951*, pp. 190-215. Oxford University Press, London.

[Firth 1957] Firth, J., 1957. A Synopsis of Linguistics Theory 1930-1955, in: *Studies in Linguistic Analysis*. Oxford Philogical Society; reprinted in: Palmer, F. (ed.), 1988. *Selected Papers of J. R. Firth*. Longman, Harlow.

[Pereira 1994] Pereira, L. A. S., 1994. *Como se combinam as palavras? Contributo para um Dicionário de Combinatórias do Português*. M.A. Thesis, Faculty of Letters, University of Lisbon, ms.

[Sinclair 1991] Sinclair, J., 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.