# Multifunctional Computational Lexicon of Contemporary Portuguese: An Available Resource for Multitype Applications

**Florbela Barreto**
**Raquel Amaro**
Centro de Linguística da Universidade de Lisboa (CLUL)
Av. Professor Gama Pinto, Nº 2
1649-003 LISBOA
{florbela.barreto; ramaro}@clul.ul.pt

**Abstract**

This paper presents some aspects of the first Portuguese frequency lexicon extracted from a corpus of large dimensions. The Multifunctional Computational Lexicon of Contemporary Portuguese (henceforth MCL) rised from the necessity of filling a gap existent in the studies of the contemporary Portuguese. Until recently, the frequency lexicons of Portuguese were of very small dimensions, such as *Português Fundamental*, which is constituted by 2.217 words extracted from a 700.000 word corpus and the *Frequency Dictionary of Portuguese Words* based on a literary corpus of 500.000 words. We describe here the main steps taken for collecting the lexical and frequency data and some of the major problems that arouse in the process. The resulting lexicon is a freely available reliable resource for several types of applications.

## 1. Introduction

MCL is a 26.443 lemma frequency lexicon with 140.315 different wordforms, with the minimum lemma frequency of 6. Each lemma and its wordforms (inflected forms and some compounds) are followed by morphosyntactic and quantitative information.

This lexicon was the result of the project *Léxico Multifuncional Computorizado do Português Contemporâneo*, financed by PRAXIS XXI programme (PRAXIS XXI/2/2.1/CSH/ 759/95), which ended in 2000.

## 2. CORLEX

MCL was extracted from a 16.210.438 word corpus - which we named CORLEX.

CORLEX is a subcorpus of *Reference Corpus of Contemporary Portuguese* (Cf. http://www.clul.ul.pt/ english/sectores/projecto_crpc.html) and contains written and spoken texts of several types; according to the international principles and recommendations established for the dimension and design of linguistic corpora meant for extracting lexica (Zampolli, 1995):

- spoken: 856.195 words; this subcorpus contains orthographic transcriptions of informal conversations and more formal productions like conferences, interviews in the radio and TV, etc;
- written subcorpus (press, literary, techno-scientific, didactic and varia): 15.354.243 words.

In order to represent the common language and a great diversity of themes, CORLEX is mainly constituted by journalistic texts, as we can observe in the figure below:

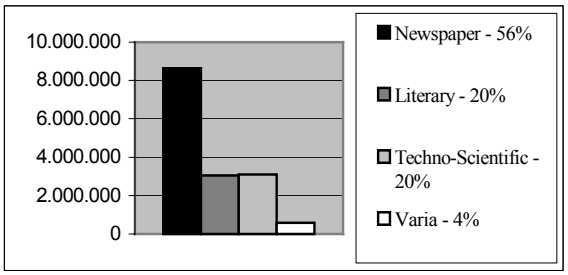


Figure 1: Genre Distribution

## 3. The Lexicon

### 3.1 Extraction of the lexicon

In order to extract the lexicon from CORLEX, all different lexical forms occurring in the corpus were indexed. Then, all wordforms were automatically tagged (morphosyntactic tagging) and lemmatised by an automatic analyser. The tags were theoretically attributed to each wordform that occurred in the corpus, i.e. the forms received all the possible tags.

The next task consisted on a manual verification of the tags attributed to each wordform and lemma (with a ocurrence frequency equal or superior to 6).

The criteria followed in this verification was the same used in *Português Fundamental* (Bacelar do Nascimento et al., 1987a, pp. 358-391).

This manual verification was a very important task since it allowed us to correct several problems:

**1.** Several wordforms were disregarded by the automatic lemmatiser and tagger, and were included manually, namely:

*a)* acronyms: *irs* (IRS), *frelimo* (FRELIMO), *palop* (PALOP);

*b)* foreign words: *homepage* (homepage), *motard* (motard);

*c)* abreviations: *dra* (Dr), *sra* (Mrs);

*d)* non-conventional orthographies: *êle* (him), *sêde* (thirst);

*e)* adverbs in *-mente*: *comprovadamente* (justifiably), *merecidamente* (worthly);

*f)* other cases of recently frequent wordforms: *clonagem* (cloning), *metadona* (methadon), *pedófilo* (pedophile).

**2.** Several lemma, which were the product of overgeneration of the automatic tagger, were withdrawn, mainly verbs, resulting in non-existent forms such as *aromar* (fragrance + Verb termination), *aurorar* (dawn + Verb termination), *bruxar* (witch + Verb termination), *cavaleirar* (knight + Verb termination), etc.

**3.** Some homographic cases ignored by the lemmatiser were considered, for instance *apoiante* (supporter) was

only considered an adjective when it can also be tagged as noun.

**4.** Other cases were incorrectly considered homographic, such as *adega* (wine cellar) that was tagged as noun and verb and after the manual verification kept only the noun tag.

After this manual verification, another lemmatisation was carried out having the following results:
Lemma with a frequency superior to 6 ---------- 30.806
Different wordforms ---------- 131.433
Homographic wordforms ---------- 44.773

### 3.2 Disambiguation

The disambiguation of homographic wordforms was made following different criteria:
Probabilistic calculations and automatic extraction of rules were made using the PAROLE annotated subcorpus[1]. We also used Eric Brill's Tagger (http://www.cs.jhu.edu/~brill/) over CORLEX for the automatic disambiguation.
In a parallel process, it was performed a manual disambiguation in order to account for the differences of: tagging between the PAROLE corpus and CORLEX, on one hand, and to reassure some *a priori* odd cases, on the other.

**1.** Differences between PAROLE and CORLEX tagging:
*a)* Ambiguous wordforms that occurred in CORLEX and didn't in the PAROLE tagged corpus: *acrescente* (add) (adjective, noun, verb), *vindima* (vintage) (noun, verb), *capricho* (caprice) (noun, verb);
*b)* Some wordforms had less categories in the PAROLE corpus than in the CORLEX. For instance the wordform *fora* (outside, out, was, went) was tagged as adverb and verb in the PAROLE corpus and adverb, interjection, noun, adposition, verb (*ser* (be) and *ir* (go)) and locution element;
*c)* Some wordforms had the same POS in the PAROLE corpus and in CORLEX. However, in the latter the wordform belongs to different lemma. For example, *revista* (magazine, reviewd, searched) is a noun and a verb in the PAROLE corpus and in CORLEX is a noun and a verb, belonging to the lemma *rever* (review) and *revistar* (search).

**2.** Other cases:
*a)* Manual validation and analysis of wordforms whose frequency and/or grammatical category seemed odd: *encantado* (delighted, charmed) had two categories, adjective and verb, and it turned out to be only adjective;
*b)* Disambiguation of wordforms with the same grammatical category but belonging to different lemma: *consumo* (consumption, consummation), for instance, is a verb and belongs to two different lemma - *consumir* (consume) and *consumar* (consummate).
After the gathering of all the mentioned data, resulting from the automatic disambiguation and from the manual disambiguation and validation, the final indexation of the Lexicon was made.

### 3.3 Morphosyntactic classification

The lemma and wordforms are marked with codes that correspond to part of speech categorisations, and other cases, as shown below:

Noun----------N
Verb----------V
Adjective----------A
Pronoun and Adjunct Pronoun----------P
Article----------T
Adverb----------R
Adposition----------S
Conjunction----------C
Numeral----------M
Interjection----------I
Foreign word----------Xf
Abreviation----------Xa
Acronym/Sigla----------Xy
Symbol----------Xs
Mediopassive *Se*----------U
Locution element----------L
Emphatic particle----------E
Displaced element----------d
Non-conventional orthography----------*
Contraction----------+
Lemma----------@
Reconstructed lemma----------[ ]
Reconstructed wordform----------<>

The following table describes the lemma and frequency distribution per POS:

| Category | Nº of different lemma | Frequency in CORLEX [2] |
|---|---|---|
| N | 14.515 | 4.115.080 |
| V | 4.154 | 2.417.516 |
| A | 6.284 | 1.060.376 |
| P | 112 | 1.203.029 |
| T | 4 | 2.534.805 |
| 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 |

Table 1: Number of occurring lemma in each of the considered categories

[1] This subcorpus contains 250.000 annotated wordforms and is included in the Portuguese corpus, as part of the project *Preparatory Action for Linguistics Resources Organisation for Language Engineering* (PAROLE - http://www.linglink.lu/le/projects/le-parole).

[2] Some wordforms may have counted more than once since they can belong to different lemma, for instance *da* (Adposition + Article) with a frequency of 231.356 occurs under the lemma *de* and *a*).

Also, wordforms with non-canonical orthography were included in their rightful lemma.

### 3.4 Quantitative Information

The quantitative information resulted from two different processes, according to the tasks described above. Each wordform frequency considers the ocurrence of a given wordform with a specific POS tag, being the lemma frequency the sum of its ocurring forms frequencies. This process was applied to the wordforms and correspondent lemma that did not pose any processing difficulties.

However, for the cases that required some more analysys and validation, we used probabilistic calculations, based on the data from the manually revised PAROLE subcorpus, in order to determine the final frequencies regarding wordforms and lemma that occured more than 200 times, and manual disambiguation frequency results for the wordforms that occurred to 200 times in the CORLEX corpus.

Therefore, the quantitative data regarding the lemma in the Lexicon (with frequency value equal or higher than 6), resulted of these calculations and of the manual disambiguation process performed.

The number of occurrences is presented along with each lemma entry and with each wordform. The frequency information is presented differently depending on the final two available formats of the lexicon. In txt format the frequencies are given by the number following the lemma and wordforms. In pdf format, since the occurrence variation interval is very wide, concerning both lemma and wordforms, a logarithmic scale, of base 10 (log10/2), was used to obtain a homogeneous distribution of the quantitative data. The data are represented by sequences of characters that indicate the value intervals, as shown below:

**Lemma:**

6 - 10 ◩□□□□□
11 - 31 ■□□□□□
32 - 100 ■◩□□□□
101 - 316 ■■□□□□
317 - 1.000 ■■◩□□□
1.001 - 3.162 ■■■□□□
3.163 - 10.000 ■■■◩□□
10.001 - 31.622 ■■■■□□
31.623 - 100.000 ■■■■◩□
100.001 - 316.227 ■■■■■□
316.228 - 1.000.000 ■■■■■◩
1.000.001 - 3.162.277 ■■■■■■

**Wordforms:**

1 - 5 ○○○○○○
6 - 10 ◐○○○○○
11 - 31 ●○○○○○
32 - 100 ●◐○○○○
101 - 316 ●●○○○○
317 - 1.000 ●●◐○○○
1.001 - 3.162 ●●●○○○
3.163 - 10.000 ●●●◐○○
10.001 - 31.622 ●●●●○○
31.623 - 100.000 ●●●●◐○
100.001 - 316.227 ●●●●●○
316.228 - 1.000.000 ●●●●●◐

## 4. Availability

MCL is presented in two different file formats, wether it is aiming mainly at consultation (.pdf files) or evaluation purposes (.txt files):

MCL is available on-line for consultation and downloading at:
www.clul.ul.pt/english/sectores/projecto_lmcpc.html

**MCL sample by alphabetical order:**

**Pdf File**

@ maçã (N) (*apple*) ■■□□□□
maçã (N) ●●○○○○
maçãs (N) ●●○○○○
maçãzinhas (N) ○○○○○○

@ macabro (A) (*macabre*) ■◩□□□□
macabra (A) ●○○○○○
macabras (A) ◐○○○○○
macabro (A) ●○○○○○
macabros (A) ○○○○○○

@ macaco (A) (*ape like, ≅ tricky*) ◩□□□□□
macaco (A) ◐○○○○○

@ macaco (N) (*ape*) ■■□□□□
macaca (N) ◐○○○○○
macacas (N) ○○○○○○
macaco (N) ●◐○○○○
macacos (N) ●◐○○○○
macaquinha (N) ○○○○○○
macaquinho (N) ○○○○○○
macaquinhos (N) ○○○○○○
macaquitos (N) ○○○○○○

**Txt File**

@ maçã (N) # 298 (*apple*)
maçã (N) # 168
maçãs (N) # 129
maçãzinhas (N) # 1

@ macabro (A) # 75 (*macabre*)
macabra (A) # 21
macabras (A) # 8
macabro (A) # 42
macabros (A) # 4

@ macaco (A) # 7 (*ape like, ≅ tricky*)
macaco (A) # 7

@ macaco (N) # 156 (*ape*)
macaca (N) # 8
macacas (N) # 1
macaco (N) # 70
macacos (N) # 69
macaquinha (N) # 1
macaquinho (N) # 4
macaquinhos (N) # 2
macaquitos (N) # 1

## 5. Conclusion

The construction of a frequency lexicon of this type and dimension can be briefly described as constituted by two major tasks:

**1.** Corpus design and constitution, according to the goals set for the final product; in our case a frequency lexicon representing general contemporary Portuguese.

**2.** Lexicon extraction, which included

- indexation of all the different lexical forms in the corpus;
- POS tagging and lemmatisation;
- wordforms frequency extraction according to each wordform POS category;
- tag and lemma assignment verification;
- manual validation and disambiguation;
- correction of the wordforms/POS frequency and lemma distribution;
- lemma frequency value computation;
- final formating.

The final lexicon, due to the quantitive and qualitative information it comprises and to the manual linguistic analysys it reflects, constitutes an important resource for studies and processing applications on contemporary Portuguese. Its dimension assures a wide coverage of the language and the frequency and POS information, as well as the lemmatisation it provides, are of great assistance in the improvement of natural language processing tools such as morphological and syntatic taggers or automatic translation tools. Its final presentation formats allow a friendly use as a reference guide for teaching and study uses, in graphic presentation (pdf format), but also allow direct and simple manipulation for automatic tools (txt format).

## References


Bacelar do Nascimento, M. F. (2001). "Um novo léxico de frequências do português". In Volume de Homenagem ao Professor Herculano de Carvalho (forthcoming).

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

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

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

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

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


## Acknowledgements


We would like to thank all the Corpus Linguistics Group of CLUL and all the people at INESC involved in this project. In special, we would like to thank Professor Maria Fernanda Bacelar do Nascimento for the extraordinary scientific guidance of this work and for supporting the writing of this paper.