TR 2008/003 ISSN 0874-338X

Frequency Analysis of the Portuguese Language

Pedro Quaresma
Department of Mathematics
University of Coimbra, Portugal

Centre for Informatics and Systems of the University of Coimbra

Frequency Analysis of the Portuguese Language

Pedro Quaresma[1]
Department of Mathematics
University of Coimbra
3001-454 COIMBRA, PORTUGAL
e-mail: pedro@mat.uc.pt   phone: +351-239 791 170

July, 2008

[1] This work was partially supported by programme POSC.

Abstract

The study of a language statistics it is very important for the cryptanalysis of substitution and/or permutation ciphers. In that type of ciphers one letter is substituted by another one, or its order is changed, with the order of another letter also from the text. In either cases the "personality" of the letter remains intact, hidden inside a different vest, but intact anyway.

If it is true that the modern block ciphers hide those characteristics, given the fact that they operate at bit level, we think that it is still important to have at hand such a tool for our own language, we can think it more has an education tool, in order to present and/or study the classical ciphers, or also has one more tool in our cryptanalyst toolbox.

In this research report we present the language statistics for the modern Portuguese language, we have analysed a large and significant set of texts, using the Portuguese alphabet, i.e. we have included in the roman alphabet the accented words and the "c" with a cedilla, and we decided to make the study case-insensitive. We present the frequency of the letters, digrams, trigrams, first letters, last letters, average length of the words, short words, and also the index of coincidence.

Keywords: Frequency analysis; Cryptanalysis.

Chapter 1

Introduction

The relative frequencies of the letters, digrams, trigrams, the first, and last, letters of a word, the average length of words, and the frequencies of the "small" words, are all characteristics of a given language [2, 3, 5, 6]. The behaviour of the letters and words reflects the way a people use its own language, and characterise that language in an unique way. Using this fact the knowledge of the different data about a language allows the cryptanalyst of substitution and/or permutation ciphers to do a comparative study, between the values found on encrypted messages, and the values given in this study, breaking, in this way, the cipher. Although the modern ciphers no longer work on letters, but on bits, we think that frequency values for a given language it is still an important tool in the cryptanalyst toolbox.

In this research report we present the frequency analysis for all the important parameters of the Portuguese language, that is, the relative frequencies of the letters in the Portuguese alphabet, the relative frequencies of digrams, trigrams, first letters, last letters, the average length of the words in the Portuguese language and the relative frequencies of the "small" words. For this we have analysed a large and significant set of texts from known Portuguese and Brazilian authors, adding in the total more then eleven millions letters, and more then two millions words.

We present bar charts with all the most important data. The full set of data is presented (in Portuguese) in `http://www.mat.uc.pt/~pedro/cientificos/Cripto/`.

This research report is organised as follows: first, in Chapter 2, we present the alphabet used in this study and we make some considerations about the text used as a base for the study of the frequencies analysis. Next, in Chapter 3, we present the most significant results in bar charts. In Chapter 4, we show, by way of two examples, how we can used the data present in order to criptoanalyse the substitution ciphers. The conclusions are given in Chapter 5. In the two appendixes we present the list of authors and web repositories used.

Chapter 2

Language Statistics for the Portuguese Language

2.1 The Texts

To study the language statistics for a given language we have to choose a large and significant set of texts, one that can represent the chosen language faithfully.

The chosen set must satisfy two main criteria:

- dimension;
- the type of texts chosen.

The texts must be of various types, covering many different authors, historical context, and kind. If someone chose to study the language statistics for a given language in a given period (e.g. modern time vs. ancient time), the authors/texts select must belong to that period [1].

Given the fact the we are interested in the modern Portuguese syntax, we have select many different "recent" authors (see appendix A). In the total we have analysed 141 texts, from 47 authors, with a counting of 11.133.372 letters and 2.400.295 words.

2.2 The Alphabet

In choosing the alphabet we decided to have a full representation of the Portuguese syntax, e.g. given the fact that in a computer encoding the accented letters and the "c" with the cedilla are distinct characters (as opposed to composed symbols), we have choose an alphabet based on the ISO 8859-1 (ISO Latin1) encoding, extending in that way the normal roman alphabet used in the Portuguese language with all the characters needed to represent the accented letters and the "c" with cedilla (see Table 2.1).

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	à	á	â	ã
ç	è	é	ê	ì	í	ò	ó	ô	õ	ù	ú	ü		

Table 2.1: Portuguese Alphabet

2.3 The Programs

To process the texts we gather them in one unique text, collating all the texts top to bottom, for that purpose we used a simple application of the Unix command `cat`. The programs needed for processing that text extracting from it the necessary information where specified in the *Flex* language[1], a program that accept a specification of a finite automata and produces the correspondent lexical analyser as a `C` program. All those programs are available at `http://www.mat.uc.pt/~pedro/cientificos/Cripto/`.

[1] `http://www.gnu.org/software/flex/manual/`

Chapter 3

The Results

In the next section we present a partial view of the obtained results. Only the most significant results are showned, the reasons for that are: space, and significance. From some point forward the difference between results, e.g. different digrams, is almost irrelevant. The complete set of results is available in *OpenOffice* spreadsheet format. They are available at the already mentioned Web page.

3.1 Letters Relative Frequencies

In the following bar chart (see Figure 3.1) we show the letters relative frequencies, for values above 0.36. As expected the “a” has the greatest value. As shown in the bar char the “ã”, “ç”, “á”, and “é” are already present, stating its importance in the Portuguese language.

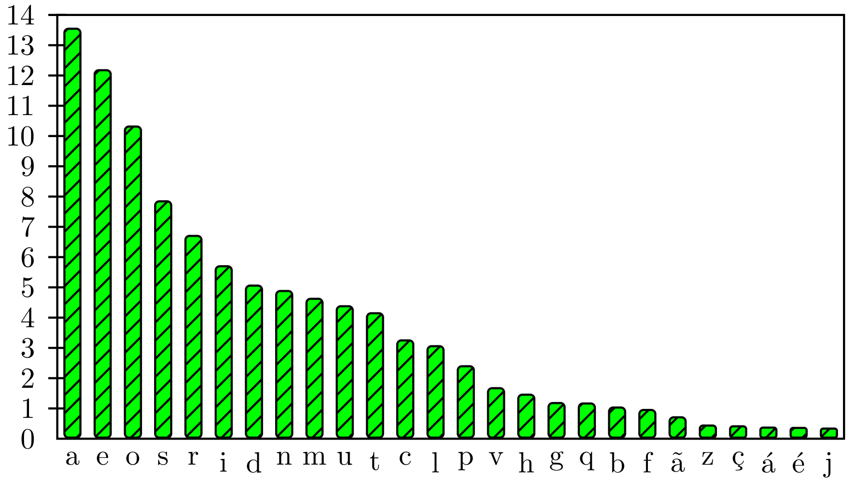

Figure 3.1: Letters Relative Frequency

3.2 Average Length of Words

In the Portuguese language the hyphen is used to break (as usual) a word between two lines, but also for some composite words like “fim-de-semana” (weekend). We questioned our colleges from the Faculty of Humanities of the University of Coimbra, the Portuguese Studies Department, on: how to count those composite words, as one, or as many? We reach a state of no conclusion. There are not a consensus about whether a given composite word should be counted as one or as n words... given this “no conclusion” status, we have decided to count the composite words as one word.

Given the fact that the hyphen is not to be counted as a letter (consensual fact) we have 2.400.295 analysed words and 11.133.372 letters given a average length of 4.638 for a word in the Portuguese language.

3.3 Short Words

Given the fact that, as said above, the average length of a Portuguese word is 4.638, we have taken as short words, the words with length one, two, and three.

3.3.1 One Letter Words

Some of the one letter words found were not actual words but contractions, e.g. “D.” for “Dona” (≈ lady), “V.” in the context of “V. Ex^{a}” (≈ Your Excellency), and others. If we look into the bar chart we can see that the actual one letter words are the only ones with a significant relative frequency value.

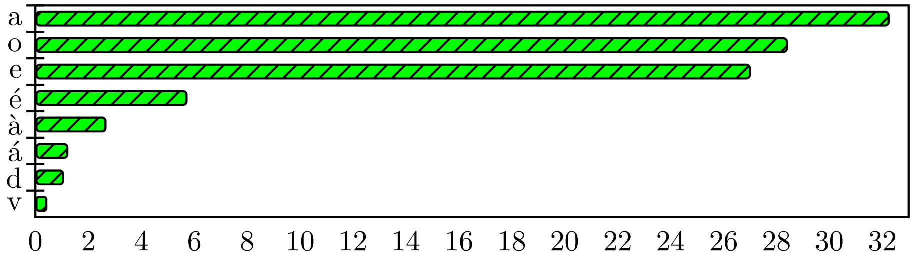

Figure 3.2: One Letter Word

3.3.2 Two & Three Letters Words

In the figures 3.3 and 3.4 we can see the relative frequencies of the most important two and three letter words in the Portuguese language.

As in the “one letter word” counting we have here some non-words,

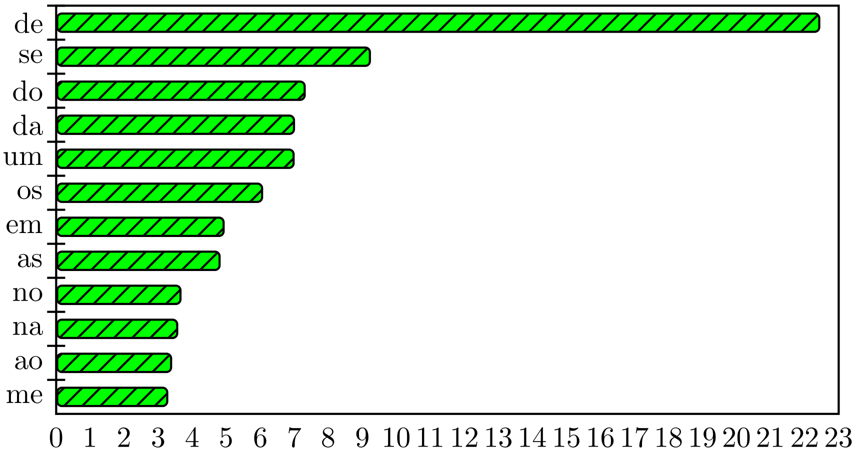

Figure 3.3: Two Letters Words

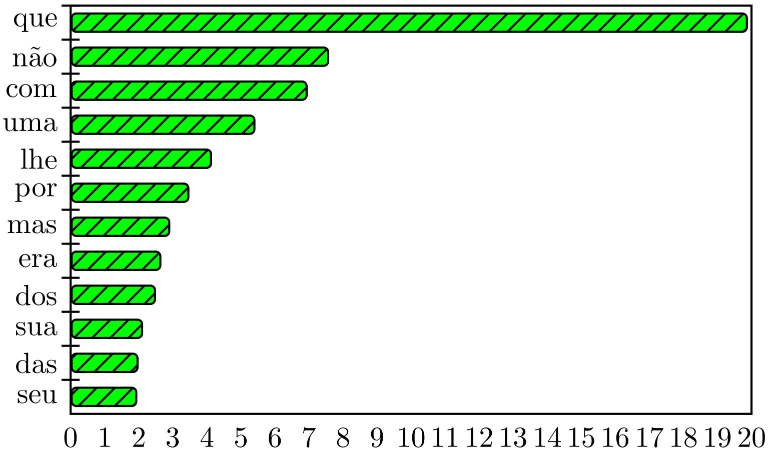

Figure 3.4: Three Letters Words

e.g. “Sr.” (≈ Sir). As in the preceding case this non-words do not have a significant impact in the overall study and can be ignored.

3.4 Digrams and Trigrams

Digrams are sequences of two letters in (any part of) a word, that is, we count any group of two letters that are part of a word. For example, *word* have the following digrams: *wo*; *or*; *rd*. The trigrams are the three letters sequences.

Digrams and Trigrams give an account of the “neighbours” that a given letter has in the Portuguese language. It is common sense that we will not have many “cx” occurring, but we will have “de”. We can see that the last one is indeed the most common digram and we can add that the first one did not occur even once in all the texts.

We have analysed 1061 different digrams and 8940 trigrams in the two following bar charts the most significant results are pictured (see Figures 3.5 and 3.6).

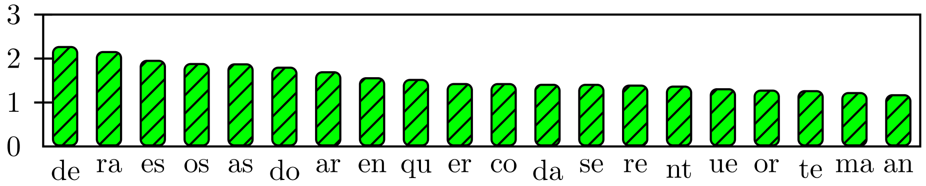

Figure 3.5: Digrams

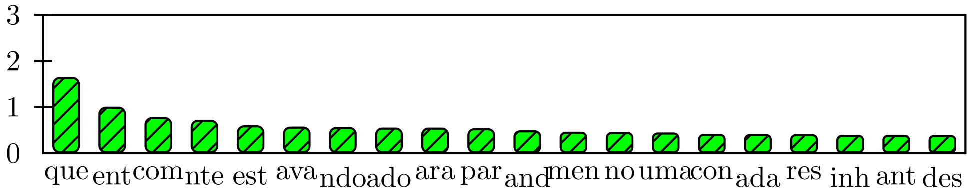

Figure 3.6: Trigrams

3.5 Initial and Final Letters

The relative frequencies of the letters that can be in the beginning, and in the end, of a Portuguese world was also studied, the results can be seen in the following figures (see Figures 3.7 and 3.8).

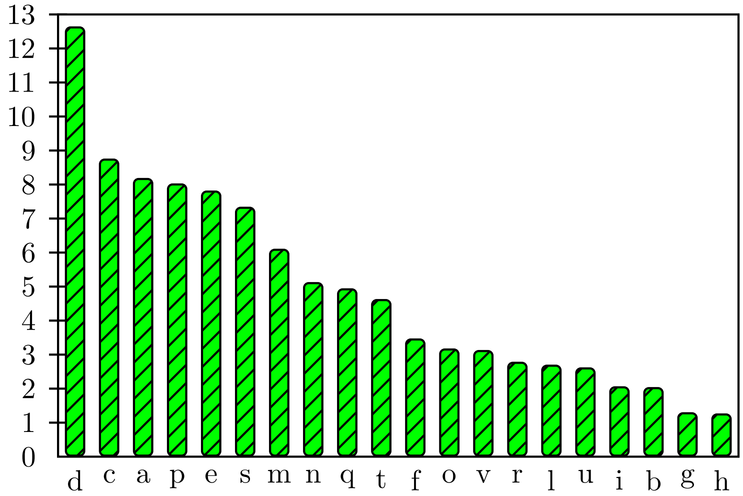

Figure 3.7: Initial Letters of Words

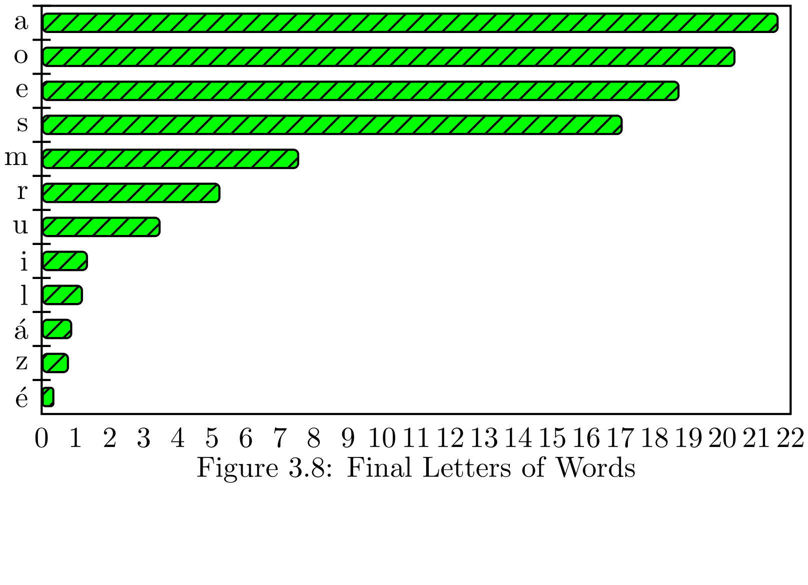

Figure 3.8: Final Letters of Words

3.6 Index of Coincidence

If for a monoalphabetic cipher the above results are enough, for a polyalphabetic cipher they are insufficient. In a polyalphabetic cipher the same character can be encrypted on many different forms, so the above results cannot be applied, first we have to find the length of the key.

The key length can be obtained by the index of coincidence. This concept was defined by Willian Friedman in 1920, as follows [2, 4]

Definition 1 (Index of Coincidence) *Suppose $X = x_1 x_2 \dots x_n$ is a string of n alphabetic characters belonging to $\mathcal{A}$. The* Index of Coincidence *of X, denoted $I_c(X)$, is defined as the probability that two random elements of X are identical.*

$$I_c(X) = \frac{\sum_{i=1}^{|\mathcal{A}|} \binom{f_i}{2}}{\binom{n}{2}} = \frac{\sum_{i=1}^{|\mathcal{A}|} f_i(f_i - 1)}{n(n-1)}$$

where f_i is the frequency of the character of $|\mathcal{A}|$ with encoding i.

Denoting the expected probabilities of occurrence of the letters of the Portuguese language in Figure 3.1 by $p_1, \dots, p_{|\mathcal{A}|}$, respectively, we have:

$$I_c(Pt) = \sum_{i_1}^{|\mathcal{A}|} p_i^2 = 0.072723$$

We can expect that for a given string of Portuguese language X its index of coincidence will be approximately equal to the value of $I_c(Pt)$, and this fact can be used to find the key length.

Chapter 4

Criptoanalysis of Substitution Ciphers

The substitution ciphers are block ciphers with symmetric keys implementing a permutation of the alphabet. This type of ciphers work substituting one character by another one, in the case of monoalphabetic substitution ciphers, or by more then one, in the case of polyalphabetic substitution ciphers [2, 4].

4.1 Criptoanalysis of a Monoalphabetic Cipher

In a monoalphabetic cipher each character of the message in encrypted as a different character, but each time the same character is encrypted in the same way. So an “a” can be disguised as a “x”, but its “DNA” remains intact. To criptoanalyse such a cipher it is enough to proceed with a frequencies analysis of the encrypted message and to compare the results with the values that we have for the language used in the message. This is a ciphertext-only attack, made by a passive adversary [2].

> m1olr fívdu r pdlv lpsruwdqwh lpshudgru urpdqr ghvorfdyd
> dv ohwudv gd phqvdjhp ruljlqdo wuòv srvlhv sdud hylwdu txh
> r lqlpljr oò-vh rv vhxv sodqrv.

Making the frequency analysis of the encrypted text we have.

Letters the most common letters were:

Letter	d	v	r
Frequency	13.0	11.4	10.6

Making the matching between this values and the data we have got for the Portuguese language we have as candidate keys: $ck_1 = \{3, 42, 32, 21, 17, 7, 13\}$.

Letter	a	e	o
d	3	42	32
v	21	17	7
r	17	13	3

We could just test all the keys in this first set, but for now we will proceed with the other measures we presented above.

Digrams In the text we have as most frequent, the following digrams: “dq”, “du”, “lp”, “ru”, “rv”, “ud”, all with a frequency $> 3\%$

	de	ra	os	es
dq	(0,12)	(29,16)	(32,41)	(42,41)
du	(0,16)	(29,20)	(32,2)	(1,41)
lp	(8,11)	(37,15)	(**40**,**40**)	(40,3)
ru	(14,16)	(0,20)	(3,2)	(13,2)
rv	(14,17)	(0,21)	(**3**,**3**)	(13,3)
ud	(17,42)	(**3**,**3**)	(6,28)	(16,28)

looking only to the pair with the same candidate key, we have $ck_2 = \{3, 40\}$.

Trigrams In the text we have as most frequent, the following trigrams: dqr, lps, , all with a frequency $> 2,7\%$.

	que	ent	com
dpr	(30,38,13)	(42,2,41)	(1,1,5)
lps	(38,38,14)	(7,2,42)	(9,1,6)

We do not get any useful information from here.

One letter words We have a single, one letter word, in the text: r.

Letter	a	e	o
r	17	13	3

so $ck_3 = \{13, 17, 3\}$.

Two letters words The text contain the following two letters words: dv, gd, oò, rv, vh.

	de	ra	os	es
dv	(0,17)	(29,21)	(32,3)	(42,3)
gd	(3,42)	(32,3)	(35,28)	(2,28)
oò	(11,32)	(40,36)	(0,18)	(10,18)
rv	(14,17)	(0,21)	(**3**,**3**)	(13,3)
vh	(18,3)	(4,7)	(7,32)	(17,32)

if we look only to the matching pairs we have: $ck_4 = \{3\}$.

Three letters words We have a single, three letters word, in the text: txh.

	que	ent	com
txh	(**3**,**3**,**3**)	(15,10,31)	(17,9,38)

In this case we have a matching triplet, so $ck_5 = \{3\}$.

The key We have seen that there is a strong evidence that the key used was $K = 3$. If we try that key in the encrypted text we get.

> júlio césar o mais importante imperador romano deslocava as letras da mensagem original três posições para evitar que o inimigo lê-se os seus planos.

We have breaked the cipher.

4.2 Criptoanalysis of a Polyalphabetic Cipher

In a polyalphabetic cipher as the Vigenère cipher [2, 3, 4, 5], a given letter will be substituted by other n different letters, where n is the length of the key chosen by the user of the cipher.

To break a cipher like the Vigenère cipher, we have to, first, find the key length, and only then we can try to find the key itself.

Given the fact that this is a block cipher, i.e. the encryption function uses the same key, over and over, until the end of the message, we know that the same sub-key is used on every n-position, of the text. We can use this fact, and the index of coincidence to find the key length.

The cripto-attack it is still a ciphertext-only attack, made by a passive adversary. Given a ciphertext we proceed by dividing it in 2 sub-texts, 3 sub-texts, ..., n sub-texts, each of which is built by the characters of the original ciphertext 2-positions, 3-positions, ..., n-positions apart, respectively.

For example, given the ciphertext "lpsruwdqwh", we have for a key length of 3, the following sub-texts: "lrdh"; "puq"; "sww". If this is the actual key length, then each of this sub-texts will have an index of coincidence with a value close of the value found for the Portuguese language. If not, the values will be quit different and we can conclude that this is not the actual key length.

Having found the key length it is easy to use the mutual index of coincidence to find the key itself [2, 4].

Chapter 5

Conclusion

We have presented in this work a frequency analysis for the "modern" Portuguese language, presenting in a graphical manner all the relevant data obtained. In appendix we present the list of authors and texts used in the study (see Appendix A). The complete tables with all values, and the programs used to produce then can be accessed in `http://www.mat.uc.pt/~pedro/cientificos/Cripto/` (in Portuguese).

Appendix A

List of texts and Authors

- Almada Negreiros (1893–1970): *A Engomadeira.*
- Agustina Bessa-Luís (1922–): *Os Amantes Aprovados.*
- Alexandre Herculano (1810–1877): *A Abóbada; A Morte do Lidador; Dama Pé-de-Cabra: Rimance de um Jogral;*
- Alfredo Campos (1847–1906): *A Filha do Cabinda.*
- Amadeu Lopes Sabino (1943–): *Clara Eugénia e as Metáforas.*
- Aquilino Ribeiro (1885–1963): *A Pele do Bombo; O Morgado de Fraião; O Pão-de-Ló; O Professor Intemerato e a Gaitinha do Capador; Os Ladrões das Almas; Tem Bom Corpo... Trabalhe!*
- Armando da Silva Carvalho (1938–): *Nome de Flor.*
- Augusto Abelaira (1926–2003): *O Arquimortes.*
- Brito Camacho (1862–1934): *O Compadre Rabino.*
- Camilo Castelo Branco (1825–1890): *Amor de Perdição; O Arrependimento; O Degredado; O Filho Natural; A Gratidão; Lagrimas Abençoadas; Novelas Do Minho; A Queda d'um Anjo; Salve Rei; Scenas Contemporaneas; A Senhora Rattazzi*
- Conde de Ficalho (1837–1903): *A Caçada do Malhadeiro; Uma Eleição Perdida; A Maluca D´a-Dos-Corvos.*
- Eça De Queirós (1845–1900): *José Matias; No Moinho; Um Poeta Lírico; Singularidades de uma Rapariga Loira; O Suave Milagre!; Os Contos; O Conde d'Abranhos; O Crime Do Padre Amaro; O Mandarim (1880); O Primo Basílio; A Cidade e as Serras; A Ilustre Casa de Ramires; A Relíquia; As Minas de Salomão; Os Maias.*

- Fernando Cabral Martins (1950–): *Aileron; Tempo a Perder.*

- Fernando Pessoa (1888–1935): *Navegar é Preciso; Poesias Inéditas; Poemas de Ricardo Reis; Poemas de Álvaro De Campos; O Guardador de Rebanhos; Poemas Inconjuntos; Mensagem; O Banqueiro Anarquista; Do Livro do Desassossego; Cancioneiro; O Pastor Amoroso; Ficções do Interlúdio/3, Para Além do Outro Oceano; O Eu Profundo de Outros Eus.*

- Fernando Venâncio (1944–): *O Romance Perdido.*

- Fialho d'Almeida (1857–1911): *Aves Migradoras; História de dois Patifes; A Idéa da Comadre Mónica; A Ruiva; Sempre Amigos; O Tio da América.*

- Florbela Espanca (1894–1930): *Livro de Máguas.*

- Gonçalo M. Tavares (1970–): *O Medo de George Steiner; Tentar Não Morrer; O Vaso.*

- Hélia Correia (1949–): *Vilegiatura.*

- Jacinto Lucas Pires (1974–): *L.*

- Jaime Rocha (1949–): *A Mulher que Aprendeu a Chorar.*

- Jorge De Sena (1919–1978): *Choro de Criança; Homenagem ao Papagaio Verde; As Ites e o Regulamento; Super Flumina Babylonis.*

- José Eduardo Agualusa (1960–): *O Homem da Luz.*

- José Martins Garcia (1941–2002): *Performance.*

- José Régio (1901–1969): *História de Rosa Brava.*

- José Rodrigues Miguéis (1901–1980): *A Chegada.*

- Júlio Dantas (1876–1962): *O Moleiro de Sula; Os Serenins de Queluz.*

- Júlio Dinis (1839–1871): *Os Fidalgos da Casa Mourisca; Uma Família Ingleza.*

- Lídia Jorge (1946–): *Leão Velho; Marido.*

- Luísa Costa Gomes (1954–): *A Cama de Pregos; Da Escada; Império do Amor.*

- Machado De Assis (1839–1908): *A Cartomante; A Causa Secreta; A Chinela Turca; A Desejada das Gentes; A Ela; A Herança; A Igreja do Diabo; A Inglezinha Barcelos; A Mão e a Luva; A Mulher de Preto; A Parasita Azul; A Segunda Vida; A Senhora do Galvão; A Sereníssima República; Adão e Eva; Americanas; Anedota do Cabriolet; Anedota Pecuniária; As Bodas de Luís Duarte; Aurora sem Dia.*

- Manuel De Arriaga (1840–1917): *Cantos Sagrados.*

- Manuel Teixeira Gomes (1860–1941): *Uma Cena Grega; Uma Copejada de Atum; D. Joaquina Eustáquia Simões d'Aljezur; O Sítio da Mulher Morta.*

- Maria Teresa Horta (1937–): *Uriel.*

- Maria Velho da Costa (1938–): *O Amante do Crato; Um Amor de Cão.*

- Mário Beirão (1890–1965): *Cintra.*

- Mário de Carvalho (1944–): *O Celacanto; A Inaudita Guerra da Avenida Gago Coutinho; A Pele do Judeu; Que Todos Ficassem Bem. . . .*

- Mário Cláudio (1941–): *Se Tu Viesses Ver-Me Hoje À Tardinha.*

- Mário Henrique Leiria (1923–1980): *A Sombra.*

- Nuno Júdice (1949–): *O Azar dos Távoras.*

- Ramalho Ortigão (1836–1915): *A Primeira Tempestade.*

- Raul Brandão (1867–1930): *Os Pobres.*

- Teixeira de Pascoaes (1877–1952): *O Doido e a Morte; Elegia da Solidão; À Ventura.*

- Teresa Veiga (1945–): *Confidência Barreirense.*

- Trindade Coelho (1861–1908): *Manhã Bendita; Manuel Maçores; Os Meus Amores.*

- Vergílio Ferreira (1916–1996): *A Galinha.*

- Vitorino Nemésio (1901–1978): *A Casa Fechada.*

Appendix B

Web repositories

The main Web repositories used are listed below, as a direct link, or as a starting point to other Web pages, those where the sources for the texts used in this work.

- Biblioteca online do conto: http://www.ficcoes.org/biblioteca_conto/index.html
- The Project Gutenberg: http://www.gutenberg.org/wiki/Main_Page
- Biblioteca Virtual do Estudante da Língua Portuguesa: http://www.bibvirt.futuro.usp.br/
- Fundação Biblioteca Nacional: http://www.cervantesvirtual.com/portal/fbn/cat_titulos.shtml
- Página sobre Eça de Queirós: http://figaro.fis.uc.pt/queiros/eca_intro.html

Bibliography

[1] Geraldo Barbosa. Pequena análise estatística da língua portuguesa: Machado de Assis e Pero Vaz de Caminha. `http://www.linguateca.pt/Repositorio/Barbosa2006.pdf`, 2006.

[2] Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone. *Handbook of Applied Cryptography.* CRC Press, 5th edition, 2001.

[3] Richard Spillman. *Classical and Contemporary Cryptology.* Prentice Hall, 2005.

[4] Douglas Stinson *Cryptography: Theory and Practice* CRC, 2006.

[5] Viktoria Tkotz. *CRIPTOGRAFIA - Segredos Embalados para Viagem.* NOVATEC Editora, São Paulo, Brasil, 2005.

[6] Viktoria Tkotz *Frequência de ocorrência de letras no Português* `http://www.numaboa.com/criptografia/criptoanalise/310-Frequencia-no-Portugues`, 8/2008.