Frequency Analysis of the Portuguese Language

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

July, 2008

\textsuperscript{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.
### Table 2.1: Portuguese Alphabet

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
<th>k</th>
<th>l</th>
<th>m</th>
<th>n</th>
<th>o</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>q</td>
<td>r</td>
<td>s</td>
<td>t</td>
<td>u</td>
<td>v</td>
<td>w</td>
<td>x</td>
<td>y</td>
<td>z</td>
<td>à</td>
<td>á</td>
<td>ã</td>
<td>ä</td>
</tr>
<tr>
<td>ç</td>
<td>ê</td>
<td>é</td>
<td>í</td>
<td>õ</td>
<td>ó</td>
<td>ô</td>
<td>õ</td>
<td>õ</td>
<td>õ</td>
<td>õ</td>
<td>õ</td>
<td>õ</td>
<td>õ</td>
<td>õ</td>
</tr>
</tbody>
</table>

#### 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/](http://www.mat.uc.pt/~pedro/cientificos/Cripto/).

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. Exa” (≈ 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](image)

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,
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](image1)

![Figure 3.6: Trigrams](image2)

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

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

where $f_i$ is the frequency of the character of $|A|$ with encoding $i$.

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

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

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.
<table>
<thead>
<tr>
<th>Letter</th>
<th>a</th>
<th>e</th>
<th>o</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>3</td>
<td>42</td>
<td>32</td>
</tr>
<tr>
<td>v</td>
<td>21</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>r</td>
<td>17</td>
<td>13</td>
<td>3</td>
</tr>
</tbody>
</table>

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%.

<table>
<thead>
<tr>
<th></th>
<th>de</th>
<th>ra</th>
<th>os</th>
<th>es</th>
</tr>
</thead>
<tbody>
<tr>
<td>dq</td>
<td>(0,12)</td>
<td>(29,16)</td>
<td>(32,41)</td>
<td>(42,41)</td>
</tr>
<tr>
<td>du</td>
<td>(0,16)</td>
<td>(29,20)</td>
<td>(32,2)</td>
<td>(1,41)</td>
</tr>
<tr>
<td>lp</td>
<td>(8,11)</td>
<td>(37,15)</td>
<td>(40,40)</td>
<td>(40,3)</td>
</tr>
<tr>
<td>ru</td>
<td>(14,16)</td>
<td>(0,20)</td>
<td>(3,2)</td>
<td>(13,2)</td>
</tr>
<tr>
<td>rv</td>
<td>(14,17)</td>
<td>(0,21)</td>
<td>(3,3)</td>
<td>(13,3)</td>
</tr>
<tr>
<td>ud</td>
<td>(17,42)</td>
<td>(3,3)</td>
<td>(6,28)</td>
<td>(16,28)</td>
</tr>
</tbody>
</table>

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%.

<table>
<thead>
<tr>
<th></th>
<th>que</th>
<th>ent</th>
<th>com</th>
</tr>
</thead>
<tbody>
<tr>
<td>dpr</td>
<td>(30,38,13)</td>
<td>(42,2,41)</td>
<td>(1,1,5)</td>
</tr>
<tr>
<td>lps</td>
<td>(38,38,14)</td>
<td>(7,2,42)</td>
<td>(9,1,6)</td>
</tr>
</tbody>
</table>

We do not get any useful information from here.

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

<table>
<thead>
<tr>
<th>Letter</th>
<th>a</th>
<th>e</th>
<th>o</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>17</td>
<td>13</td>
<td>3</td>
</tr>
</tbody>
</table>

so \( ck_3 = \{13, 17, 3\} \).

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

<table>
<thead>
<tr>
<th></th>
<th>de</th>
<th>ra</th>
<th>os</th>
<th>es</th>
</tr>
</thead>
<tbody>
<tr>
<td>dv</td>
<td>(0,17)</td>
<td>(29,21)</td>
<td>(32,3)</td>
<td>(42,3)</td>
</tr>
<tr>
<td>gd</td>
<td>(3,42)</td>
<td>(32,3)</td>
<td>(35,28)</td>
<td>(2,28)</td>
</tr>
<tr>
<td>oò</td>
<td>(11,32)</td>
<td>(40,36)</td>
<td>(0,18)</td>
<td>(10,18)</td>
</tr>
<tr>
<td>rv</td>
<td>(14,17)</td>
<td>(0,21)</td>
<td>(3,3)</td>
<td>(13,3)</td>
</tr>
<tr>
<td>vh</td>
<td>(18,3)</td>
<td>(4,7)</td>
<td>(7,32)</td>
<td>(17,32)</td>
</tr>
</tbody>
</table>

if we look only to the matching pairs we have: \( ck_4 = \{3\} \).
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 them can be accessed in http://www.mat.uc.pt/~pedro/cientificos/Cripto/ (in Portuguese).
• 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 Inconjurados; Mensagem; O Banqueiro Anarquista; Do Livro do Desassossego; Caucioneiro; 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 Idéia da Comadre Mônica; A Ruiva; Sempre Amigos; O Tio da América.

• Florbela Espanca (1894–1930): Livro de Mágicas.

• 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 Aprende a Chorar.

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

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


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


• Júlio Dantas (1876–1962): O Moleiro de Sula; Os Serenins de Queiauz.

• 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.
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/bibliotecaconto/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/
- Página sobre Eça de Queirós: http://figaro.fis.uc.pt/queiros/eca_intro.html