Análise de Frequências da Língua Portuguesa

Pedro Quaresma e Augusto Pinho

Resumo — Na actual Aldeia Global da Idade da Informação o estudo/compreensão da criptografia e da cripto-análise (criptologia) afigura-se como muito importante. Nesse contexto desenvolveu-se uma página da Rede que apresenta as cifras mono-alfabéticas de substituição clássicas, permitindo a experimentação das mesmas, assim como o explorar dos métodos cripto-analíticos para as "quebrar", a saber o estudo de frequências e a procura exaustiva no espaço das chaves. Neste artigo apresenta-se o estudo estatístico da Língua Portuguesa, nos que diz respeito às suas principais características. É este estudo que suporta os métodos cripto-analíticos para as cifras clássicas, as quais devido ao seu carácter de cifras de substituição mantêm inalteradas as características próprias de uma língua. Apresentam-se de forma gráfica os resultados mais importantes para: a frequência relativa das letras, os digramas, os trigramas, as letras iniciais, as letras finais, e as "palavras curtas".

Palavras-Chave — Criptografia, Crito-análise, Língua Portuguesa Contemporânea.

1 Introdução

A frequência das letras; dos digramas; dos trigramas; das primeiras e últimas letras de uma palavra; das "palavras curtas"; assim como o comprimento médio das palavras, são características de uma dada língua. As ocorrências das letras e das palavras reflectem a forma como um povo utiliza a sua língua, caracterizando-a de forma única. O cripto-analista pode então usar este conhecimento único para analisar as mensagens encriptadas através de cifras de substituição.

As cifras de substituição, mono-alfabéticas ou poli-alfabéticas, caracterizam-se por encriptar um texto claro através da substituição dos caracteres que aí surgem por outros, de forma única (cifras mono-alfabéticas), ou sob várias transformações (cifras poli-alfabéticas). No entanto para ambos os casos mantêm-se as propriedades dos diferentes caracteres [1], [2], [3].

Com o advento das cifras modernas (DES, RSA, etc.) este tipo de ataque cripto-analítico deixa de ser possível, passando a sua utilidade a estar associada à aprendizagem da criptologia. Foi no âmbito do desenvolvimento de uma página da Rede dedicada à apresentação da criptologia [4], [5] que se desenvolveu o estudo da análise de frequências da Língua Portuguesa (http://www.mat.uc.pt/~pedro/cientificos/Cripto/).

- *Pedro Quaresma, Departamento de Matemática, Universidade de Coimbra, Coimbra. pedro@mat.uc.pt.*
- *Augusto Pinho: gustopinho@iol.pt...*

No que se segue apresenta-se o estudo estatístico de todos os parâmetros importantes da Língua Portuguesa, a saber: a frequência relativa das letras do alfabeto Português (estendido); a frequência relativa dos digramas, e dos trigramas; a frequência relativas das primeiras e últimas letras das palavras; o comprimento médio das palavras; e finalmente, a frequência relativa das "palavras pequenas". Para este efeito analisou-se um conjunto muito extenso e representativo de autores Portugueses contemporâneos, perfazendo um total de mais de 5 milhões de letras.

A informação mais significativa é apresentada através de gráficos de barras. A informação completa pode ser descarregada a partir da página da Rede acima referida.

2 Dados Estatísticos para a Língua Portuguesa

2.1 Os Textos escolhidos

Para fazer um estudo estatístico de uma dada língua é necessário escolher um conjunto amplo e significativo de textos, os quais devem poder representar fielmente a língua em estudo.

O conjunto escolhido deve respeitar dois critérios principais:

- Dimensão, isto é, o número de caracteres contados;
- tipo: os textos devem ser de diferentes tipos, cobrindo muitos autores, contextos históricos, e tipos literários.

Se se pretende fazer o estudo estatístico de uma dada língua num dado período (por exemplo, o estudo do Português Medieval), os textos/autores devem pertencer a esse período [6].
Dado que pretendíamos estudar o Português moderno, escolheram-se autores "recentes" (desde meados do século XIX até à actualidade), versando alguns estilos diferentes, com ênfase em romancistas. No total analisaram-se 95 textos de 38 autores, totalizando 5.112.633 caracteres e 1.102.087 palavras (ver a secção 4).

2.2 O Alfabeto

Ao escolher o alfabeto de base para o nosso estudo decidiu-se ter a representação completa do alfabeto Português, isto é, o alfabeto escolhido é o alfabeto latino estendido para comportar as letras acentuadas, assim como o 'c' cedilhado (veja-se a tabela 1).

Tabela 1
Alfabeto Português

A	B	C	D	E	F	G	H	I	J	K	L
0	1	2	3	4	5	6	7	8	9	10	11
M	N	O	P	Q	R	S	T	U	V	W	X
12	13	14	15	16	17	18	19	20	21	22	23
Y	Z	a	b	c	d	e	f	g	h	i	j
24	25	26	27	28	29	30	31	32	33	34	35
k	l	m	n	o	p	q	r	s	t	u	v
36	37	38	39	40	41	42	43	44	45	46	47
w	x	y	z	À	Á	Â	Ã	Ç	È	É	Ê
48	49	50	51	52	53	54	55	56	57	58	59
Ì	Í	Î	Ñ	Ò	Ó	Ô	Õ	Ù	Ú	Û	Ü
60	61	62	63	64	65	66	67	68	69	70	71
à	á	â	ã	ç	è	é	ê	ì	í	î	ñ
72	73	74	75	76	77	78	79	80	81	82	83
ò	ó	ô	õ	ù	ú	ü					
84	85	86	87	88	89	90					

3 Apresentação dos Resultados

3.1 Frequência Relativa das Letras

Para uma cifra de substituição mono-alfabética aditiva, caracterizada pela função de cifração $e_c(x) = (x + c) \bmod 91$ com *c* a chave secreta, e 91 o comprimento do alfabeto considerado, a cripto-análise não é um tarefa muito complicada.

É fácil de ver que um ataque por procura exaustiva no espaço das chaves (só há 90 chaves possíveis), usualmente referido como um ataque de "força bruta", é praticável, podemos dizer mesmo que é trivial se se considerar o uso de um computador.

No entanto se se optar por uma cifra de substituição um pouco mais "complexa", como seja a Cifra Linear, cuja função de cifração é $e_c(x) = (ax + b) \bmod 91$ com o par *(a,b)* a definir a chave secreta, então a utilização de um ataque de força bruta, já não é tão fácil de executar. O espaço das chaves tem 6552 chaves possíveis, uma tarefa já virtualmente impossível sem o recurso a um computador.

Para esta última cifra (assim como para a primeira), assim como para outras cifras de substituições mais complexas, como sejam as poli-alfabéticas, um método cripto-analítico possível, e fácil de montar, é dado pelo estudo das frequências das letras das mensagens cifradas.

Os métodos de substituição trocam uma dada letra por uma (ou mais) letras diferentes, por exemplo, na Cifra Aditiva, com chave 3 (usualmente designada por Cifra de Júlio César), tem-se que um 'a' é substituído por um 'd', um 'b' é substituído por um 'e', etc. Mas então as características próprias da letra 'a' na Língua Portuguesa (a sua frequência relativa, etc.), vão ficar inalteradas, mas agora "escondido" na letra 'd'. O estudo comparativo dos valores encontrados para a Língua Portuguesa e dos valores encontrados no(s) texto(s) da(s) mensagem(ns) cifrada(s), vai permitir obter um emparelhamento entre letras, e dessa forma obter a chave secreta, quebrando a cifra.

No gráfico que se segue (ver Figura 1), mostram-se as frequências relativas das letras para valores acima de 0,2.

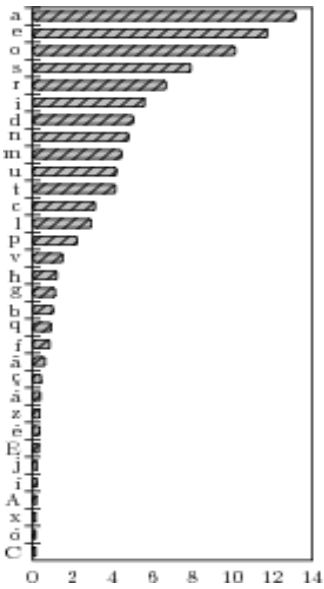

Fig. 1: Frequência Relativa das Letras

Como esperado a letra 'a' é aquela que tem um maior valor de frequência relativa. Se não se considerar as letras maiúsculas como diferentes das minúsculas (prática usual em

muitos sistemas), pode-se ver que as posições relativas não se alteram. Como se pode ver no gráfico os caracteres 'ã', 'ç', e as vogais agudas (com a excepção do 'ú'), já marcam posição, constatação evidente da sua importância na Língua Portuguesa. As primeiras letras maiúsculas são o "E", o "A", e o "C".

3.2 Comprimento Médio das Palavras

O ajuste entre as frequências relativas dos caracteres nas mensagens cifradas e na Língua Portuguesa, não é, na grande maioria dos casos, perfeito. Este facto deve-se evidentemente à menor dimensão do texto da mensagem, assim como ao tipo de texto da mensagem original. Quanto maior for o número de mensagens cifradas usadas no estudo das frequências relativas mais os valores obtidos se aproximarão dos valores de referência encontrados neste estudo. No entanto além dos valores das frequências relativas dos diferentes caracteres constituintes do alfabeto Português é ainda possível utilizar outras particularidades da Língua Portuguesa para, em conjunto, podermos fazer o estudo cripto-analítico da cifra.

Um valor que vai determinar o estudo de outros elementos a estudar é o valor do comprimento médio das palavras. Esse valor é determinante para caracterizar o conceito de "palavra curta".

A determinação do comprimento médio das palavras não deixa de levantar alguns problemas, um deles prende-se com o uso do hífen. Na Língua Portuguesa o hífen é usado para a translineação das palavras, mas também para algumas palavras compostas, como por exemplo "fim-de-semana". Após algumas consultas com colegas da Faculdade de Letras da Universidade de Coimbra, do Departamento de Estudos Portugueses, concluiu-se que não havia um consenso sobre se as palavras compostas deviam contar como um só palavra, ou se deviam contar todas as suas palavras constituintes. No estudo realizado optou-se por contar as palavras compostas como uma única palavra. Não se chegou a fazer nenhum estudo do impacto desta decisão, que se estima não ser muito importante.

Dado o facto de que o hífen não é contado como letra (facto que reuniu consenso) obteve-se então a contagem de 1.102.087 palavras e 5.112.633 letras, dando portanto um comprimento médio de 4,64 para uma palavra em Português. Com base nesse valor definiu-se "palavra curta" como uma palavra com menos de quatro letras.

Vejamos de seguida o estudo efectuado para os restantes elementos que podem ser usados para caracterizar uma dada língua.

3.3 Digramas e Trigramas

Os digramas e os trigramas são conjuntos de duas ou três letras seguidas e que constituem sub-palavras, isto é, que fazem parte de uma palavra. Eles vão dar uma ideia das "vizinhanças" que se encontram numa dada língua, ou seja quais as letras que aparecem mais vezes associadas a outras letras, dentro das palavras.

Analisaram-se 1736 digramas diferentes e 10674 trigramas diferentes, os resultados mais relevantes são apresentados nas figuras 2 e 3.

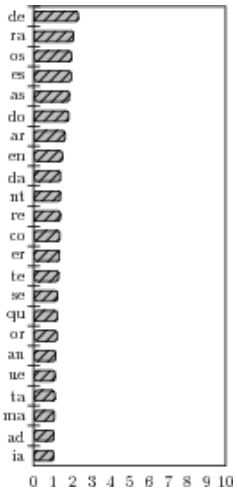

Fig. 2: Digramas

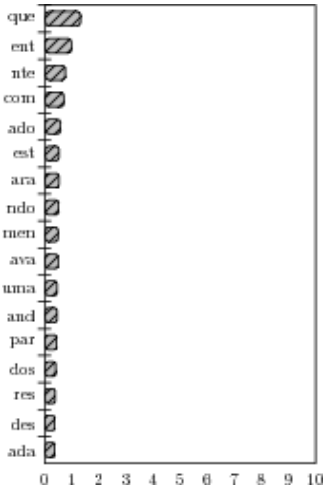

Fig. 3: Trigramas

3.4 Palavras Curtas

Dado que, como foi dito acima, a média do comprimento das palavras em Português é de 4,64 tomaram-se como "palavras curtas", isto é, palavras cujo comprimento está abaixo da média, as palavras com comprimento um, dois e três. Note-se que não se está perante uma nova contagem dos digramas e dos trigramas, neste caso só se contabilizou as palavras próprias cujo comprimento total é um dos valores em apreciação, não se contabilizaram as sub-palavras.

3.4.1 Letras Isoladas

Trata-se de contabilizar as palavras cujo comprimento é um, isto é, letras isoladas que pertencem ao léxico Português.

Algumas das palavras de uma só letra são afinal contracções de outras palavras de maior comprimento, por exemplo 'D.' (para "Dona" ou "Don"), 'V.' (para "Vossa"), assim como outras. Embora esses casos surjam no gráfico a sua expressão já não é significativa.

Na figura 4 apresentam-se os dados mais significativos obtidos para palavras de comprimento um.

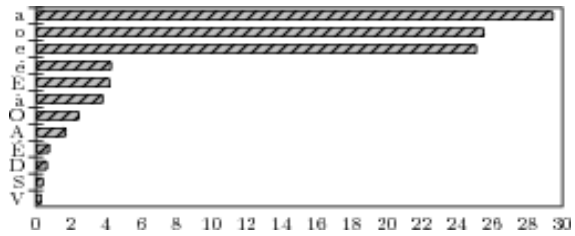

Fig. 4: Palavras de Uma Letra

3.4.2 Palavras com Duas e Três Letras

De seguida (vejam-se as figuras 5 e 6) apresentam-se os dados mais significativos para as palavras de comprimento dois e três respectivamente.

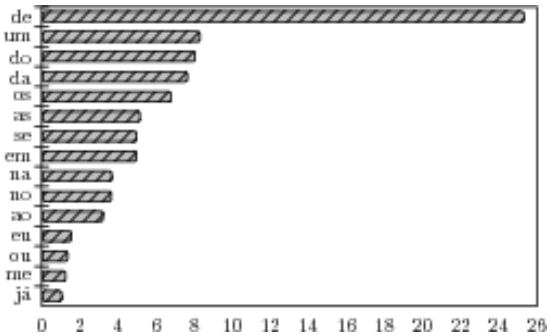

Fig. 5: Palavras de Duas Letras

Assim como para as letras isoladas surgem algumas "não-palavras" na forma de contracções, por exemplo "Sr." (Senhor), assim como alguns casos relacionados com a hifenização das palavras. O impacto de tais ocorrências não é significativo no estudo global podendo ser ignoradas.

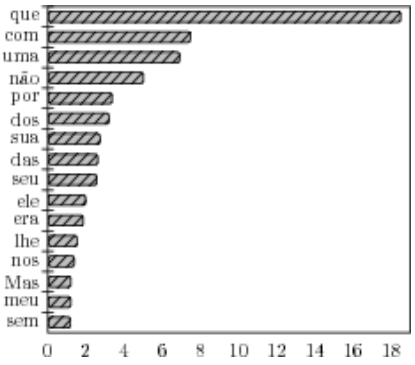

Fig. 6: Palavras de Três Letras

3.5 Letras Iniciais e Letras Finais

Um dado também importante no estudo de uma dada língua é a frequência relativa das letras que estão no começo das palavras. O mesmo se pode dizer para as letras que estão no fim das palavras.

Fez-se então o estudo estatístico das letras que podem constituir o início de uma palavra, isto é, a primeira letra de cada uma das palavras dos textos estudados. Fez-se também igual estudo para todas as possíveis letras que estão na última posição das palavras.

Os resultados mais significativos são apresentados nas figuras 7 e 8.

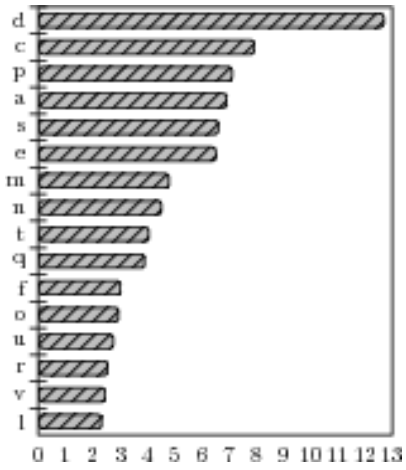

Fig. 7: Letras Iniciais de Palavras

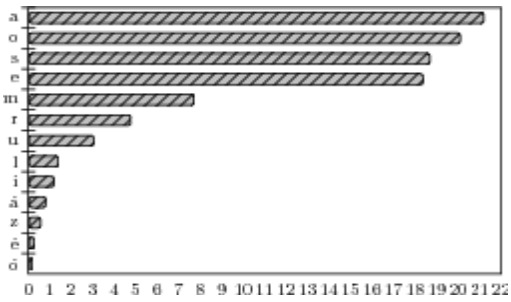

Fig. 8: Letras Finais de Palavras

4 Lista de Autores e Obras Consultadas

De seguida apresenta-se a tabela dos autores, e respectivas obras consultadas. Todos os textos estavam disponíveis electronicamente em http://www.ficcoes.net/biblioteca_conto/ e em http://figaro.fis.uc.pt/queiros/eca_intro.html.

Tabela 2: Lista de Autores e Obras

Agustina Bessa-Luís (1922-)	Os Amantes Aprovados.
Alexandre Herculano (1810-1877)	A Abóbada; A Dama Pé-de-Cabra; A Morte do Lidador.
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 Silva Carvalho (1938-)	Nome de Flor.
Augusto Abelaira (1926-2003)	O Arquimortes.
Brito Camacho (1862-1934)	O Compadre Rabino.
Camilo Castelo Branco (1825-1890)	Gracejos que Matam; Maria Moisés; O Comendador; O Degredado; O Filho Natural.
Conde de Ficalho (1837-1903)	A Caçada do Malhadeiro; A Maluca d'A-dos-Corvos; Uma Eleição Perdida.
Eça de Queirós (1845-1900)	A Capital; A Cidade e as Serras; A Correspondência de Fradique Mendes; A Ilustre Casa de Ramires; A Relíquia; Alves e Ca.; As Minas de Salomão (tradução); Cartas de Inglaterra; Crónicas de Londres; Cartas Familiares e Bilhetes de Paris; Contos; Ecos de Paris; O Conde d'Abranhos (e "a Catástrofe"); O Crime do Padre Amaro; O Mandarim; O Mistério da Estrada de Sintra; O Primo Basílio; Os Maias; Prosas Bárbaras; Últimas Páginas; Uma Campanha Alegre I; Uma Campanha Alegre II; Civilização.
Fernando Cabral Martins (1950-)	Aileron; Tempo a Perder.
Fernando Venâncio (1944-)	O Romance Perdido.
Fialho d'Almeida (1857-1911)	A Idéa da Comadre Mónica; A Ruiva; História de Dois Patifes; O Tio da América; Sempre Amigos.
Gonçalo M. Tavares (1970-)	O Medo de George Steiner; O Vaso; Tentar Não Morrer.
Hélia Correia (1949-)	Vilegiatura.
Jacinto Lucas Pires (1974-)	L.
Jaime Rocha (1949-)	A Mulher que Aprendeu a Chorar.
José Régio (1901-1978)	História de Rosa Brava.
Jorge de Sena (1919-1978)	As Ites e o Regulamento; Choro de Criança; Homenagem ao Papagaio Verde; Super Flumina Babylonis.
José Eduardo Agualusa (1960-)	O Homem da Luz.
José Martins Garcia (1941-)	Performance.
José Rodrigues Miguéis (1901-1980)	A Chegada.
Júlio Dantas (1876-1962)	O Moleiro de Sula; Os Serenins de Queluz.
Lídia Jorge (1946-)	Leão Velho; Marido.
Luísa Costa Gomes (1954-)	A Cama de Pregos; Da Escada; Império

	do Amor.
Machado de Assis (1839-1908)	Ideias de Canário.
Manuel Teixeira Gomes (1860-1941)	Dona Joaquina Eustáquia; O Sítio da Mulher Morta; Uma Cena Grega; Uma Copejada de_Atum.
Maria Velho da Costa (1938-)	Um Amor de Cão; O Amante doCrato.
Mário Cláudio (1941-)	Se Tu Viesses Ver-me Hoje à Tardinha.
Mário de Carvalho (1944-)	A Inaudita Guerra da Avenida Gago Coutinho; A Pele do Judeu; O Celacanto; Que Todos Ficassem Bem....
Henrique Leiria (1923-1980)	A Sombra de Mário.
Nuno Júdice (1949-)	A Camponesa, a Égua e o Cavaleiro; O Azar dos Távoras.
Ramalho Ortigão (1836-1915)	A Primeira Tempestade.
Teresa Veiga (1945-)	Confidência Barreirense.
Trindade Coelho (1861-1901)	Os Meus Amores; Manhã Bendita; Manuel Maçores.
Vergílio Ferreira (1916-1996)	A Galinha.
Vitorino Nemésio (1901-1978)	A Casa Fechada.

5 CONCLUSÃO

Neste texto apresentaram-se de forma gráfica os resultados mais importantes de um estudo estatístico para a Língua Portuguesa Contemporânea. Os dados completos obtidos estão acessíveis na página http://www.mat.uc.pt/~pedro/cientificos/Cripto/

.

O desenvolvimento desta página, com mais métodos criptográficos e cripto-analíticos, por exemplo o índice de coincidência, é um trabalho que está em desenvolvimento, e que continuará no futuro.

AGRADECIMENTO

Este trabalho foi suportado, parcialmente, pelo programa POSC.

REFERÊNCIAS

[1] Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone. Handbook of Applied Cryptography. CRC Press, 5th edition, 2001.
[2] Richard Spillman. Classical and Contemporary Cryptology. Prentice Hall, 2005.
[3] Viktoria Tkotz. CRIPTOGRAFIA - Segredos Embalados para Viagem. NOVATEC Editora, São Paulo, Brasil, 2005.
[4] Pedro Quaresma e Elsa Lopes. Criptografia. Gazeta de Matemática, aceite para publicação.
[5] Pedro Quaresma e Augusto Pinho. Cripto-análise. Gazeta de Matemática, aceite para publicação.
[6] Geraldo Barbosa. Pequena análise estatística da língua portuguesa: Machado de Assis e Pero Vaz de Caminha. http://www.linguateca.pt/Repositorio/Barbosa2006.pdf, 2006.

Pedro Quaresma Doutoramento em Informática, área de conhecimento de Fundamentos da Computação, pela Universidade do Minho em 1998. Professor Auxiliar do Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra. Investigador do CISUC na área da Demonstração Automática de Teoremas.

Augusto Pinho Licenciatura em Matemática, ramo científico, especialização em Computação, pelo Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra em 2006. Prémio Talento 2007, Inova-Ria 2007, patrocinado pela Sony-Ericsson.