Série Edificações do Imaginário.

I. Pequena análise estatística da língua portuguesa: Machado de Assis e Pero Vaz de Caminha

Geraldo A. Barbosa

QuantaSec Ltda.– Consultoria, Pesquisa e Projetos em Criptografia Quântica e Estatística, Av. Portugal 1558, Belo Horizonte MG 31550-000, Brasil *

(Dated: 19 de janeiro de 2006)

Resumo: Este trabalho mostra a estatística de ocorrência de letras na língua portuguesa em textos de Machado de Assis, contabilizando mais de um milhão de letras. Apresenta-se também o bigrama de ocorrência de letras para a obra "Memórias Póstumas de Brás Cubas". Foram calculados os valores da entropia para vários textos de Machado e feita uma breve discussão de flutuações estatísticas de acordo com o tamanho das amostragens. De maneira similar analisa-se a "Carta" de Pero Vaz de Caminha, no original e em versão contemporânea. Algumas comparações são feitas e discutidas ao final do trabalho.

INTRODUÇÃO

Dos sons guturais à representação sonora mais elaborada das idéias estabeleceu-se em passado pré-histórico a linguagem Proto-Indo-Européia. Na evolução desta estabeleceu-se a maioria das linguagens atuais. A representação da linguagem sonora através da escrita define outro salto qualitativo e intricado da mente humana que, felizmente, deixou muitos registros preciosos para a linguística.

A complexidade da evolução de cada linguagem escrita deixa assinaturas inequívocas que permite não somente a identificação das linguagens particulares assim como suas inter-relações e até suas evoluções históricas. O advento do computador trouxe ferramentas de análise das linguagens que permitem a linguística avanços importantes que antes eram praticamente impossíveis pelo volume de dados a serem tratados.

Dentre as assinaturas mais fundamentais de cada linguagem está a ocorrência de letras e suas combinações. Para exemplificar, observemos as palavras em inglês e português "pale" e

a	0.0575
b	0.0128
c	0.0263
d	0.0285
e	0.0913
f	0.0173
g	0.0133
h	0.0313
i	0.0599
j	0.0006
k	0.0084
l	0.0335
m	0.0235
n	0.0596
o	0.0689
p	0.0192
q	0.0008
s	0.0508
s	0.0567
t	0.0706
u	0.0334
v	0.0069
w	0.0119
x	0.0073
y	0.0164
z	0.0007

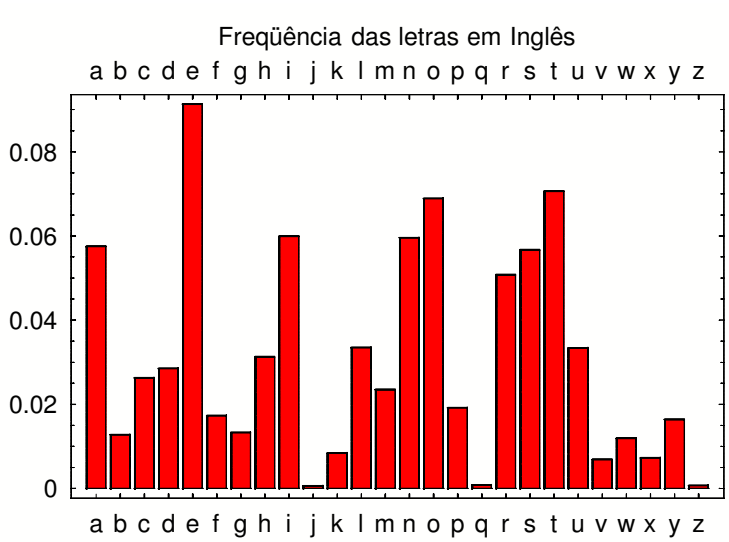


FIG. 1: Freqüência das letras na língua inglesa. A tabela à esquerda da figura mostra cada letra utilizada e sua probabilidade de ocorrência no texto.

"pela" ou "train" e "intra", ou ainda, "pious" e "pisou", "goad" e "gado", "blue" e "bule". Cada uma destas palavras, em inglês e português, foram escritas com as mesmas letras. Entretanto, não é muito fácil encontrar palavras em duas línguas escritas com as mesmas letras e em igual número. Mais difícil ainda se torna quando estas palavras contém muitas letras, embora o número de permutações de letras cresça enormemente e seja igual ao fatorial do número de letras (Para N letras, existem $N!$ permutações possíveis).

Assim, a freqüência da ocorrência das letras numa língua torna-se uma assinatura ou *impressão digital* – ou se utilizando a palavra em moda, o "DNA", daquela língua. A simples existência de símbolos especiais numa língua já impossibilita a ocorrência de palavras com as mesmas letras. Assim as letras na palavra em inglês "tea" não poderão ser rearranjadas como "até" em português, pela inexistência do acento. A ocorrência de duas letras consecutivas ou bigramas (por exemplo, "pe", "el", "la" em "pela"), trigramas e outras combinações de ordem maior são também características determinantes de cada língua (As notações bigramas e dígramas ou dígrafos, etc. são entendidas como equivalentes).

A figura 1 e tabela ao lado mostram a ocorrência (percentual) das letras em inglês retirada de uma grande amostragem de textos. Observe-se que a letra "e" é a que mais ocorre na língua inglesa, seguida de "t" e "o" e assim por diante, nas proporções indicadas.

Qual é a freqüência da ocorrência das letras em português? A resposta a esta pergunta

deve ser conhecida por muitas pessoas mas não está facilmente disponível na literatura eletrônica aberta pela Internet. A motivação para este trabalho deve-se a que o autor observou pessoas (além dele mesmo) buscando estas respostas na Internet e não terem encontrado estas estatísticas. Espera-se, portanto, que outros pesquisadores que já tenham feito estudos semelhantes sobre a língua portuguesa entendam as dificuldades para terceiros encontrarem estes trabalhos. Mais do que isto, estas dificuldades levam outros à "redescoberta da roda" e a não utilização de obras existentes. Assim, trabalhos preciosos deixam de ser compartilhados. Felizmente, isto também livra aqueles autores de quaisquer ônus nas análises aqui feitas. Estas, por sua vez, são análises comuns em estatística e linguística e podem ser reproduzidas com facilidade com "softwares" estatísticos adequados à linguagem simbólica.

Nota-se que as análises estatísticas linguísticas devem utilizar um razoável número de amostras para lhes conferirem confiabilidade como representativas daquela língua. *Qual é o número de amostras necessárias para se representar uma língua?* Este número deve ser grande e variado para não se utilizar repetições típicas de um determinado ramo do conhecimento. Uma obra como um conto genérico de maior porte talvez possa ser utilizada como representativa da língua numa determinada época. Igualmente, obras representativas de diferentes épocas podem ser utilizadas para se estudar a evolução da língua ao longo do tempo. Este trabalho pretende dar algumas respostas qualitativas e quantitativas sobre algumas destas questões.

As análises estatísticas linguísticas são tão importantes que elas fazem parte até do arsenal de grandes potências desde a Idade Média para o deciframento de mensagens dos concorrentes econômicos ou inimigos. A criptanálise tem tido papel extremamente relevante em muitos eventos de importância mundial [1]. A freqüência de textos cifrados pode, em muitos casos, revelar se um ciframento é monoalfabético, a possível freqüência ou dimensão da senha ou chave de ciframento e outras peculiaridades. A criptologia, juntamente com as técnicas de interceptação de mensagens, são parte do cotidiano de nações modernas [2]. Assim, as grandes potências conhecem segredos comerciais de nações antes de reuniões econômicas mundiais ou seus segredos militares tais como localizações de radares de outros países, grandes rêdes de transporte de energia elétrica e outras informações de importância estratégica de seu interesse. Assim, não somente os linguístas se interessam pelo estudo puro das línguas mas muitos outros vivem de seu estudo por diferentes motivações.

Este trabalho visa simplesmente fornecer um estudo de ocorrência de letras e bigra-

mas na língua portuguesa utilizando textos eletrônicos facilmente acessíveis no momento. Para uma análise estatística mais completa da língua um número bem maior e variado de textos deve ser analisado. Esta análise demanda tempo e recursos não compatíveis com o modesto objetivo do presente trabalho. Será mostrado, entretanto, que os resultados indicam convergência para índices possivelmente estacionários. Estes valores assintóticos seriam os índices representativos da língua portuguesa nesta época.

A ESCOLHA DE TEXTOS

Alguns textos foram escolhidos para análise e que pudessem mostrar alguns detalhes de diferentes épocas. Num primeiro grupo, o autor escolheu para análise alguns textos de Machado de Assis (1839/1908). Inicia-se com "Memórias Póstumas de Brás Cubas" (1881), disponível na Internet através da Fundação Biblioteca Nacional Departamento Nacional do Livro, do MEC (Ministério da Cultura – Brasil) (http://www.machadodeassis.org.br/obras003.htm). Para suporte das conclusões, utilizou-se os textos "Dom Casmurro" e "Quincas Borba", ambos digitalizados pelo Núcleo de Pesquisas em Informática, Literatura e Lingüstica http://www.cce.ufsc.br/ alckmar/literatura/literat.html, da Universidade Federal de Santa Catarina e, por último, deste mesmo autor utilizou-se os "Contos Fluminenses" (também disponibilizado pelo MEC, Brasil). O segundo grupo refere-se às versões moderna e antiga da "Carta" de Pero Vaz de Caminha (1450/1500), enviada a "El Rei D. Manuel", por ocasião da descoberta do Brasil (1500). Este texto está disponível na Internet através da Biblioteca Nacional Digital (Portugal), no enderêço eletrônico http://www.bnd.bn.pt/ed/viagens/brasil/obras/carta_pvcaminha/index.htm. As duas obras se distanciam em cerca de 400 anos. "Memórias" foi escrita em capítulos curtos e a "Carta" está dividida no original em páginas (frente e verso).

As estatísticas destes dois grupos serão apresentadas, com maior ou menor detalhe, de acordo com a necessidade da análise.

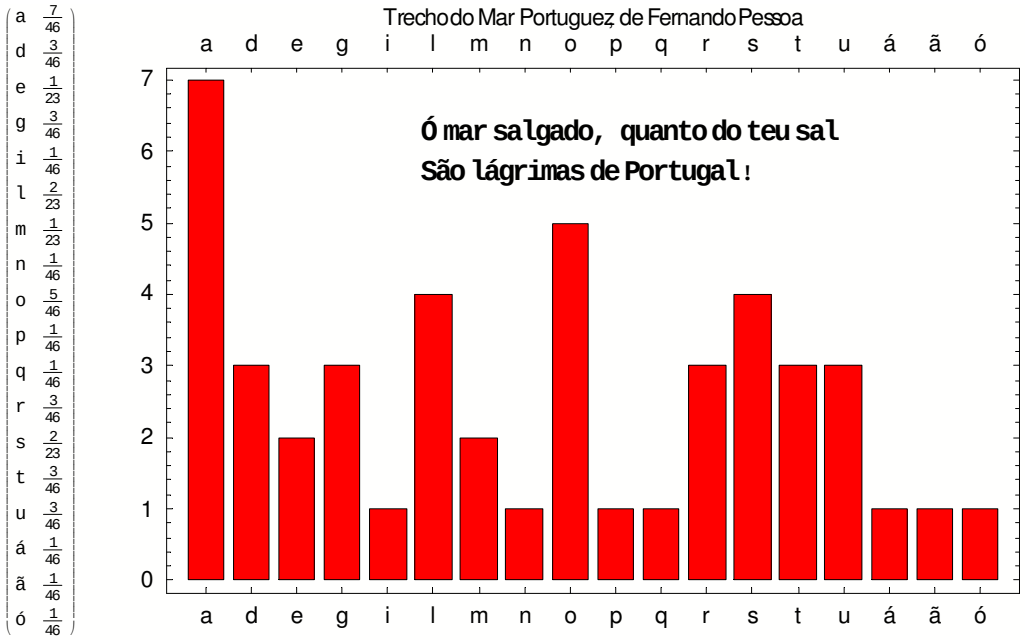


FIG. 2: Freqüência das letras em trecho de poesia de Fernando Pessoa. A tabela à esquerda da figura mostra, ao lado de cada letra, sua probabilidade de ocorrência $P(x)$ em relação ao número total de letras no texto (46 letras).

METODOLOGIA

Para as análises estatísticas realizadas foram desenvolvidos programas de computação simbólica com suporte básico do "software" *Mathematica*, versão 5, da Wolfram Research. Os resultados serão espelhados na estatística de ocorrência de letras, na ocorrência de bigramas e nas entropias dos textos analisados. A entropia de cada amostragem analisada é definida pela Teoria da Informação de Shannon [3] como

$$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)} \ , \tag{1}$$

onde x especifica cada variável e X o conjunto destas variáveis. Aqui, X designa o alfabeto usado e x as letras do mesmo. $P(x)$ é a probabilidade de ocorrência da letra x na amostragem feita, isto é, a contagem de quantos letras x ocorreram dividida pelo número total de letras na amostragem utilizada. Muitas propriedades são associadas à entropia e se constituem em parte da Teoria da Informação. A base 2 utilizada para o logarítmo é usualmente adotada e particularmente prática para a comunicação binária $(0, 1)$. As entropias resultantes são valores em "bits". Para exemplificar numericamente a construção

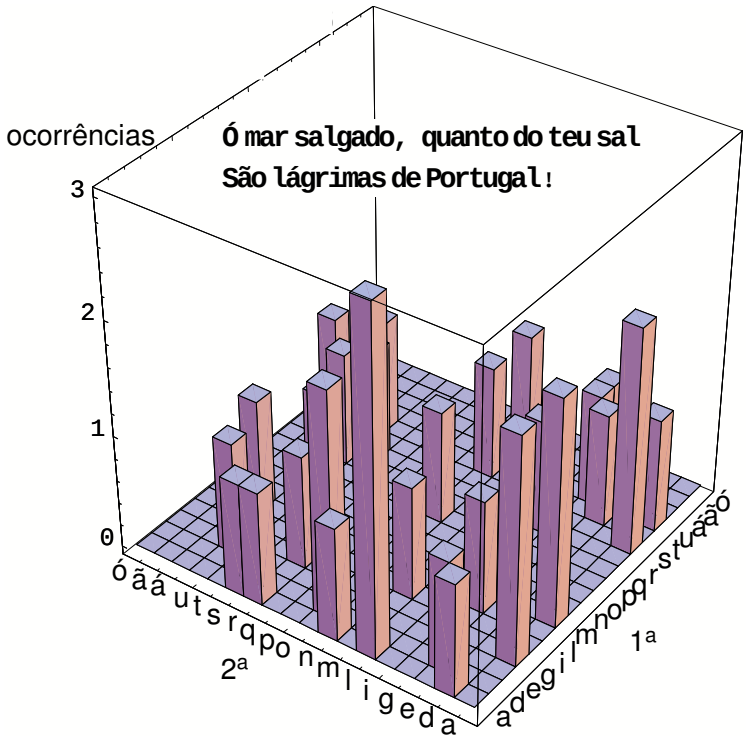


FIG. 3: Ocorrência de bigramas em trecho de poesia de Fernando Pessoa. Escolhe-se uma letra no eixo indicado 1^a e a letra seguinte no outro eixo 2^a.

das ocorrências e respectivas probabilidades, a figura 2 mostra um pequeno trecho de poesia de Fernando Pessoa com a contagem das letras. A probabilidade de cada letra está dada na lista ao lado. O mesmo trecho pode ser analisado em relação a bigramas, trigramas etc. A figura 3 mostra o bigrama do trecho escolhido. Dos bigramas pode-se retirar com facilidade as probabilidades adjuntas $P(x, y)$, assim como voltar as probabilidades $P(x)$, através de $P(x) = \sum_y P(x, y)$, etc.

MEMÓRIAS PÓSTUMAS DE BRÁS CUBAS

Freqüências

As figuras e tabelas a seguir representam a freqüência de ocorrência de letras do capítulo 1 ao capítulo 10 (Figura 4). As tabelas ao lado de cada figura representam numericamente o mesmo conteúdo da figura e são apresentadas para uso dos leitores que se interessam por aspectos númericos mais detalhados.

a	0.126906
b	0.00882916
c	0.0308134
d	0.0483299
e	0.124707
f	0.011063
g	0.0126587
h	0.0136515
i	0.0638253
j	0.00248209
k	0.0000354585
l	0.0284022
m	0.0525849
n	0.0484717
o	0.0993192
p	0.0225516
q	0.0125168
r	0.0620523
s	0.0745337
t	0.0438976
u	0.0488618
v	0.0186866
w	0.0000354585
x	0.00262393
y	0.0000354585
z	0.00450323
à	0.00152471
á	0.00485781
â	0.000425502
ã	0.00918375
ç	0.0041841
è	0.0000354585
é	0.00631161
ê	0.000992837
í	0.00382952
ó	0.00294305
ô	0.000957379
õ	0.000673711
ú	0.00141834
û	0.000070917
ü	0.000212751

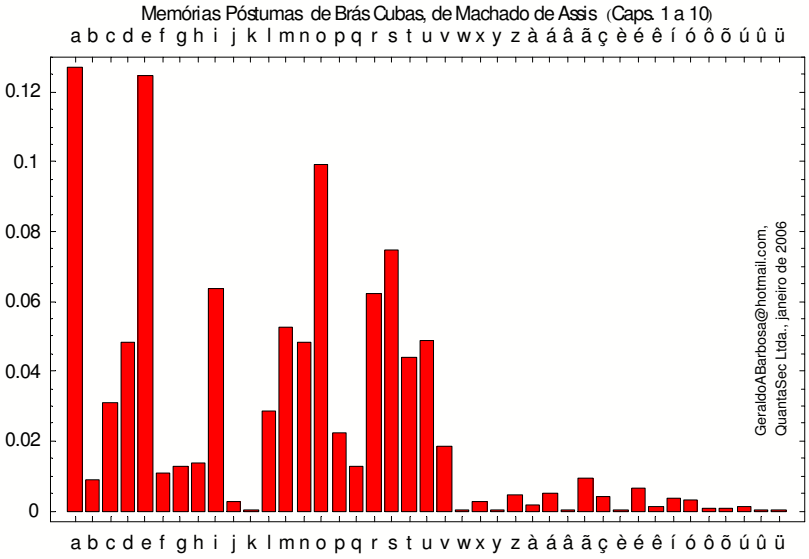


FIG. 4: Freqüência das letras nos capítulos de 1 a 10 (28.202 letras) de Memórias Póstumas de Brás Cubas. A tabela à esquerda da figura apresenta númericamente a probabilidade de cada letra.

Observa-se que as letras "a","e" e "o" são as de maior ocorrência. A letra "a" ocorre com freqüência maior do que a letra "e" do inglês. Será esta uma flutuação devida à pequena amostragem ou uma indicação de diferenças existentes entre estas línguas?. A resposta encontra-se na próxima figura. A figura 5 mostra as ocorrências de letras em todos os 160 capítulos. A maior ocorrência da letra "a" persiste assim como as freqüências são distintas daquela apresentada na Fig. 1. Será que esta amostragem pode ser vista como característica da língua portuguesa? A seção a seguir responderá a esta pergunta. De fato, já foi possível perceber que a formação de palavras numa língua é de formação complexa e com baixíssima probabilidade de ocorrência mesmo nas palavras isoladas. Um texto completo numa língua é, portanto, muito característico daquela língua. Tal raciocínio já antevê a resposta qualitativa. Entretanto, pode ser dada uma resposta quantitativa? Passemos a próxima seção.

Entropias

A figura 6 mostra as entropias dos capítulos de 1 a 10, de 1 a 20, e assim sucessivamente até o último capítulo, de número 160. A esquerda estão os valores das entropias encontradas e, à direita, estes valores divididos pelo número de letras ou símbolos do alfabeto utilizado. Observe-se que o símbolo "ñ", não pertencente à lingua portuguesa foi incluido devido a

a	0.129969
b	0.0100656
c	0.0331197
d	0.0470983
e	0.122632
f	0.0101314
g	0.0121409
h	0.0129629
i	0.0642373
j	0.00361786
k	0.0000292287
l	0.0296379
m	0.0509748
n	0.0483771
o	0.102147
p	0.0245046
q	0.0123455
r	0.0624689
s	0.0745623
t	0.0429186
u	0.0477815
v	0.0173326
w	0.0000109607
x	0.00251001
y	0.0000292287
z	0.0043368
à	0.00128241
á	0.00410663
â	0.000361705
ã	0.00959065
ç	0.00466562
è	7.30716×10^{-6}
é	0.00481542
ê	0.00167699
í	0.00418335
ñ	3.65358×10^{-6}
ó	0.00185967
ô	0.000471312
õ	0.000558998
ú	0.000906088
û	7.30716×10^{-6}
ü	0.000160758

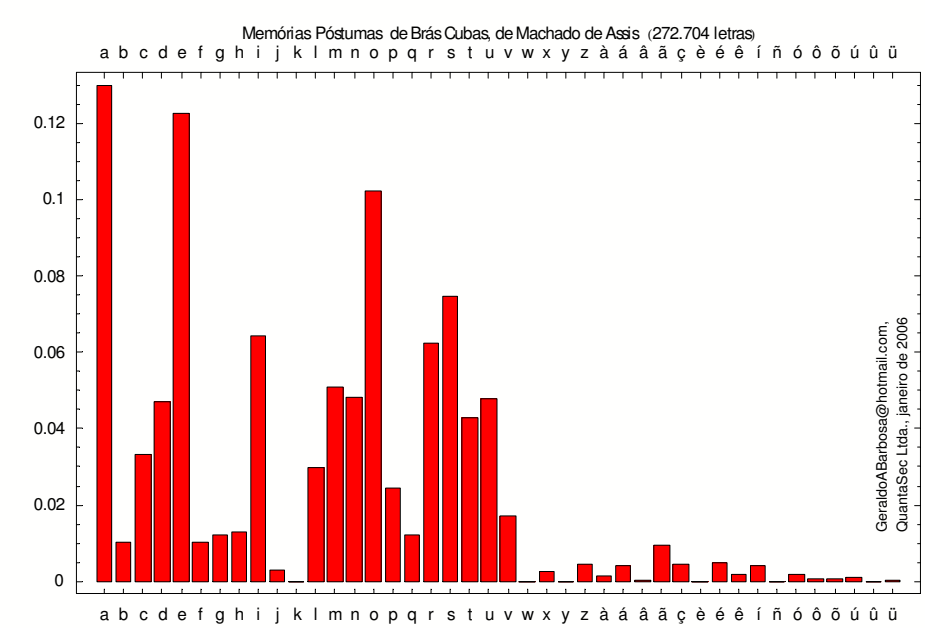


FIG. 5: Freqüência das letras nos capítulos de 1 a 160 de Memórias Póstumas de Brás Cubas. A tabela à esquerda da figura apresenta númericamente as letras encontradas e sua probabilidade de ocorrência.

citação de palavra espanhola no texto. Entretanto sua contribuição à entropia (valores à esquerda da figura) é desprezivel. Os 35 símbolos considerados no texto são 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,í,ñ,ó,ô,õ, ú, û, ü. Desprezar-se o símbolo "ñ" traria uma correção por fator multiplicativo de $34/35 \simeq 0,97$ nos valores à direita. Deve-se notar que uma amostragem reduzida produz valores da entropia que variam consideravelmente. Ao se aumentar a amostragem considerada, com o acréscimo de capítulos, observa-se que estas flutuações são reduzidas ou suavizadas. A tabela I apresenta os valores numericos das entropias obtidas. Espera-se que um limite assintótico seja atingido para números *finitos* amostrados para a entropia e a entropia por símbolo. Pode-se indagar se a amostragem utilizada já apresenta a estatística representativa da língua portuguesa. Entende-se também que pequenas flutuações poderão sempre ser notadas numa obra devido à particularidades daquele trabalho. Assim, num trabalho de caráter romântico, a palavra "amor" eventualmente poderá ocorrer com maior freqüência do que a palavra "lobo", com mesmo número de letras. Estas flutuações produzirão desvios em torno de uma média representativa da língua.

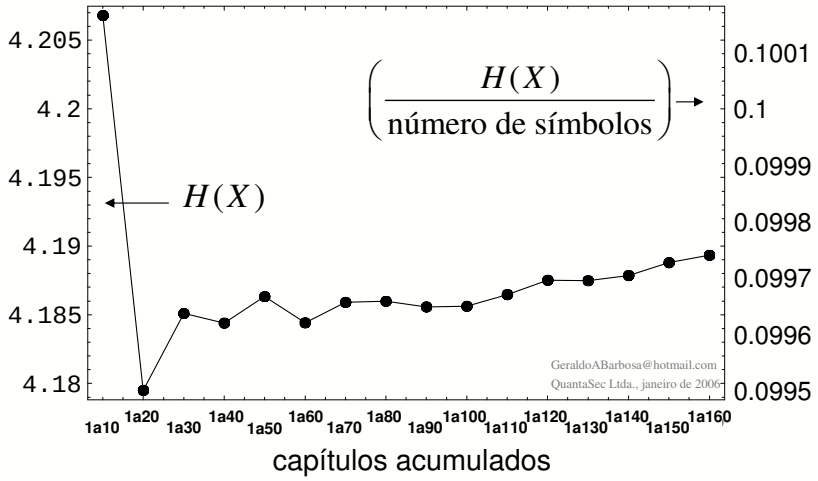


FIG. 6: Entropia de capítulos acumulados de 10 em 10. A esquerda estão os valores das entropias. À direita, a entropia por símbolo.

Redundância

Num alfabeto com L símbolos, a entropia é maximizada quando todos os símbolos têm a mesma probabilidade $P = 1/L$. Neste caso, a entropia seria (Ver Eq. 1)

$$H_{max} = \underbrace{L}_{\text{Número de símbolos}} \times \overbrace{\left(\frac{1}{L}\right)\log_2\left(\frac{1}{\frac{1}{L}}\right)}^{\text{Entropia por símbolo}} = \log_2 L \ . \tag{2}$$

Este valor máximo sugere a medida da redundância r da língua em primeira ordem, isto é, a diferença relativa entre $H(X)$ para a ocorrência de letras e H_{max}:

$$r_1 = 1 - \frac{H(X)}{H_{max}} = 1 - \frac{H(X)}{\log_2 L} \ . \tag{3}$$

Se considerarmos que o valor r_1 representa a redundância da língua (o que não é uma boa aproximação, ver abaixo), seu valor sugere que a eficiência da língua possa ser melhorada neste percentual, ou seja, de alguma forma a língua poderia ser melhorada ou compactada para ser mais eficiente. A entropia encontrada para os textos analisados forneceu $H(X) \simeq 4.19$ com um alfabeto de 35 letras. Assim, $r_1 \simeq 0,223$ ou $22,3\%$. Note que uma "mensagem" arbitrariamente longa, consistindo da repetição de um mesmo símbolo teria entropia zero e, portanto, teria a máxima redundância possível ($r = 1$). Isto significa que a mensagem, a despeito de seu tamanho, não leva qualquer informação.

Como mencionado, a entropia pode ser calculada em aproximações mais representativas: assim como a Eq. (1) foi aplicada para a probabilidade de ocorrência de letras, ela

Caps.	**Entropia**	**Entropia/símbolo**
1 a 10	4.20679	0.100162
1 a 20	4.17949	0.0995118
1 a 30	4.18509	0.099645
1 a 40	4.18439	0.0996284
1 a 50	4.18632	0.0996743
1 a 60	4.18441	0.0996289
1 a 70	4.18590	0.0996643
1 a 80	4.18598	0.0996661
1 a 90	4.18557	0.0996564
1 a 100	4.18562	0.0996575
1 a 110	4.18647	0.0996779
1 a 120	4.18751	0.0997027
1 a 130	4.18748	0.0997019
1 a 140	4.18786	0.0997109
1 a 150	4.18880	0.0997333
1 a 160	4.18933	0.0997461

TABLE I: Entropia e entropia por símbolo.

poderia também ser estendida para os bigramas, trigramas e assim por diante. Como as maiores palavras numa língua tem um número de termos relativamente pequeno ("anti-constitucionalíssimamente"?), estes cálculos precisariam ser estendidos somente até estes valores.

Sendo N_{max} o comprimento do maior $n-$grama ocorrendo na língua, em princípio poderíamos então calcular entropias em várias aproximações:

$$H_1(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}; \quad H_2(X) = \frac{1}{2} \sum_{x,y} P(x,y) \log_2 \frac{1}{P(x,y)}; \ldots$$
$$H_{N_{max}}(X) = \frac{1}{N_{max}} \sum_{\underbrace{x,y,\ldots}_{N_{max}}} P(\underbrace{x,y,\ldots}_{N_{max}}) \log_2 \frac{1}{P(\underbrace{x,y,\ldots}_{N_{max}})} \ . \tag{4}$$

As entropias em ordem mais alta fornecem valores mais representativos para a língua do

que a entropia calculada simplesmente em primeira ordem. De fato, as entropias de ordem menor podem ser calculadas a partir das de ordem maior. Por exemplo, a probabilidade para acerto de um dado símbolo β numa língua, conhecendo-se um símbolo inicial α, será dada por

$$P(\beta|\alpha) = \frac{P(\alpha,\beta)}{P(\alpha)} \ , \tag{5}$$

e assim por diante. Entretanto, os cálculos passam a demandar muito tempo. Pode-se também optar pela estatística de ocorrência das próprias palavras, sejam como unidades independentes uma das outras ou até mesmo se considerando suas ocorrências vinculadas, isto é, a probabilidade de ocorrência de uma dada palavra condicionada ao aparecimento anterior de uma ou outras palavras.

Apesar de ser uma aproximação "pobre" a entropia calculada em primeira ordem é, entretanto, uma *característica* da língua e suficiente para os objetivos deste trabalho.

Deve-se observar que Shannon [4] se utilizou também de outros procedimentos bastante pragmáticos para se calcular a redundância de uma língua, como experimentos com textos "ocultos", dos quais somente se conheciam uma certa percentagem das palavras. Por exemplo, apresenta-se o texto com um conjunto de espaços a serem preenchidos, igual ao número de letras e espaços do texto faltante. Passa-se a tentar adivinhar as letras ou espaços deste texto. O número da tentativas incorretas são anotadas para cada posição. Cada acerto é preenchido e age como uma informação para se diminuir o número de tentativas das letras ou espaços restantes. A partir destas contagem de tentativas constroem-se as probabilidades envolvidas para se acertar as letras ou espaços e, destas, constroem-se as entropias para cada letra (Um experimento deste tipo está sendo realizado atualmente através do "site" http://www.numaboa.com.br/index.php, desenvolvido por Viktoria Tkotz). Neste tipo de entropia encontram-se elementos de difícil estimativa numérica a-priori pois as probabilidades para acerto de cada símbolo sucessivo dependerá de forma correlacionada com todos elementos anteriores.

Neste trabalho, conforme utilizado na figura 6 e tabela I, as entropias por símbolo são determinadas automáticamente através das probabilidades e uso da equação 1 e da divisão do resultado pelo número de símbolos do alfabeto.

a	0.13161
b	0.00859413
c	0.0318112
d	0.0465385
e	0.12406
f	0.0097985
g	0.0109332
h	0.0134394
i	0.0611788
j	0.00334158
k	0.0000104424
l	0.0263672
m	0.0524942
n	0.0481014
o	0.101173
p	0.0267605
q	0.0147308
r	0.0609247
s	0.0790075
t	0.0416514
u	0.0479656
v	0.0170699
w	0.0000104424
x	0.0020502
y	6.96163×10^{-6}
z	0.00459816
à	0.00120784
á	0.00476524
â	0.000142713
ã	0.010982
ç	0.00449373
é	0.00621326
ê	0.00176477
í	0.00229734
ó	0.00208501
ô	0.000365486
õ	0.000549969
ú	0.000814511
ü	0.0000905012

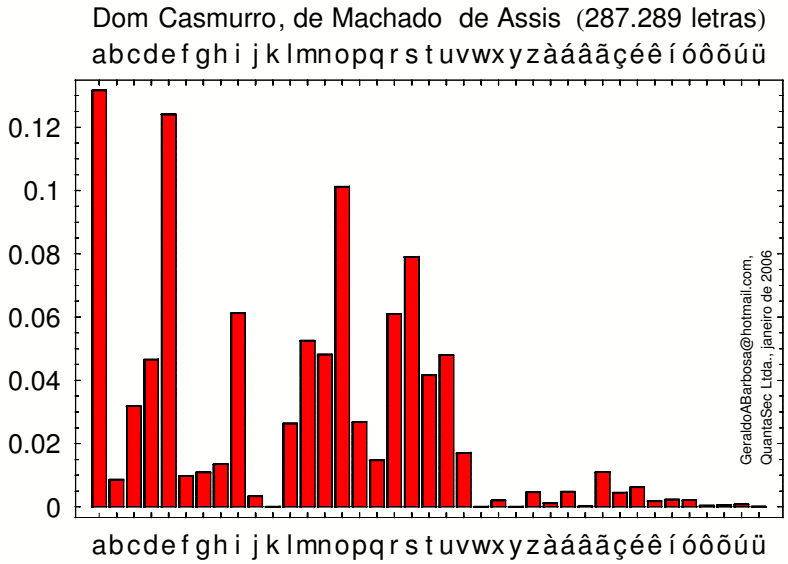


FIG. 7: Freqüência das letras em Dom Casmurro. A tabela à esquerda da figura apresenta numericamente as letras encontradas e sua probabilidade de ocorrência.

Outros textos de Machado de Assis

As figuras 7, 8, 9 mostram as freqüências em "Dom Casmurro", "Quincas Borba" e "Contos Fluminenses".

Entropias

As entropias obtidas para os textos analisados de Machado de Assis, juntamente com a contagem das letras (entre parenteses) em cada texto são: $X_{Bras\,Cubas} = 4,208$ (273.704); $X_{Dom\,Casmurro} = 4,180$ (287.289); $X_{Quincas\,Borba} = 4,173$ (342.190); $X_{Contos\,Fluminenses} = 4,167$ (251.196). O total de letras considerado é acima de um milhão de letras: 1.154.379. A entropia média obtida para os textos, de forma ponderada, é

$$\overline{H}_W = \frac{273.704X_{BC} + 287.289X_{DC} + 342.190X_{QB} + 251.196X_{CF}}{1.154.379} \simeq 4,182 \ . \qquad (6)$$

A entropia média, não ponderada, variância σ^2 e o desvio padrão σ são, respectivamente

$$\overline{H} = \frac{X_{BC} + X_{DC} + X_{QB} + X_{CF}}{4} = 4,187 \qquad (7)$$

$$\sigma^2 = 0,000289 \qquad (8)$$

a	0.134782
b	0.0120138
c	0.0320787
d	0.0487653
e	0.117525
f	0.0104708
g	0.0112218
h	0.0150852
i	0.0598352
j	0.00278208
k	0.0000175341
l	0.028528
m	0.0455712
n	0.0474824
o	0.107081
p	0.0258424
q	0.0132616
r	0.0667056
s	0.0773079
t	0.0400859
u	0.0485987
v	0.0160905
w	5.84471×10^{-6}
x	0.002183
y	8.76706×10^{-6}
z	0.00389842
à	0.00152547
á	0.0040796
â	0.000192875
ã	0.0106987
ç	0.00431924
è	0.0000116894
é	0.0045764
ê	0.00161898
ë	2.92235×10^{-6}
í	0.00204565
ó	0.00194044
ô	0.000423741
õ	0.000593238
ú	0.000625384
ü	0.000116894

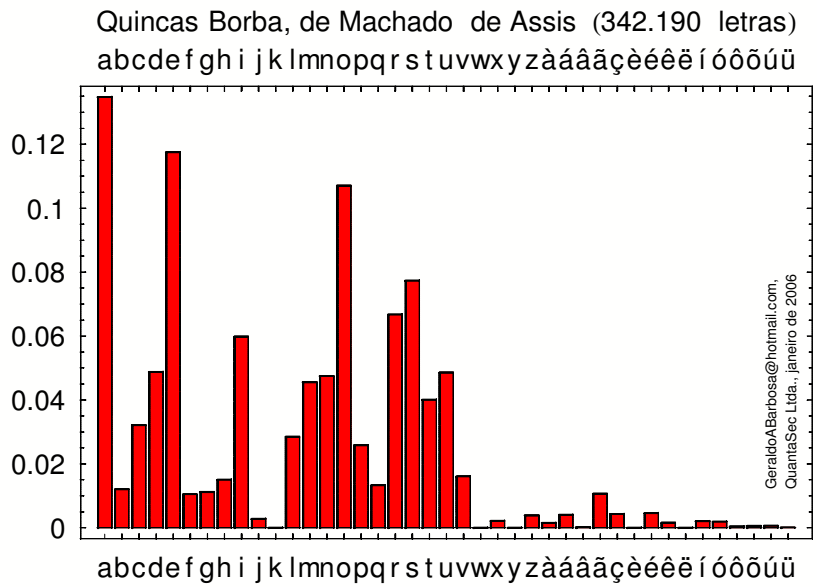


FIG. 8: Freqüência das letras em Quincas Borba. A tabela à esquerda da figura apresenta numericamente as letras encontradas e sua probabilidade de ocorrência.

a	0.12954
b	0.00690298
c	0.0313779
d	0.0506776
e	0.128469
f	0.00922387
g	0.0107964
h	0.0134278
i	0.06104
j	0.00262743
k	0.0000159238
l	0.0271103
m	0.0498137
n	0.0492245
o	0.102032
p	0.0254463
q	0.0130655
r	0.0627199
s	0.0765458
t	0.0430978
u	0.0468519
v	0.0184637
w	0.0000159238
x	0.00174764
y	0.0000278667
z	0.00519913
à	0.00134556
á	0.00387745
â	0.0003344
ã	0.00972945
æ	3.98096×10^{-6}
ç	0.00524292
è	3.98096×10^{-6}
é	0.00534244
ê	0.00206612
í	0.00305737
ó	0.00132168
ô	0.000485677
õ	0.000601124
ú	0.000983296
ü	0.000143314

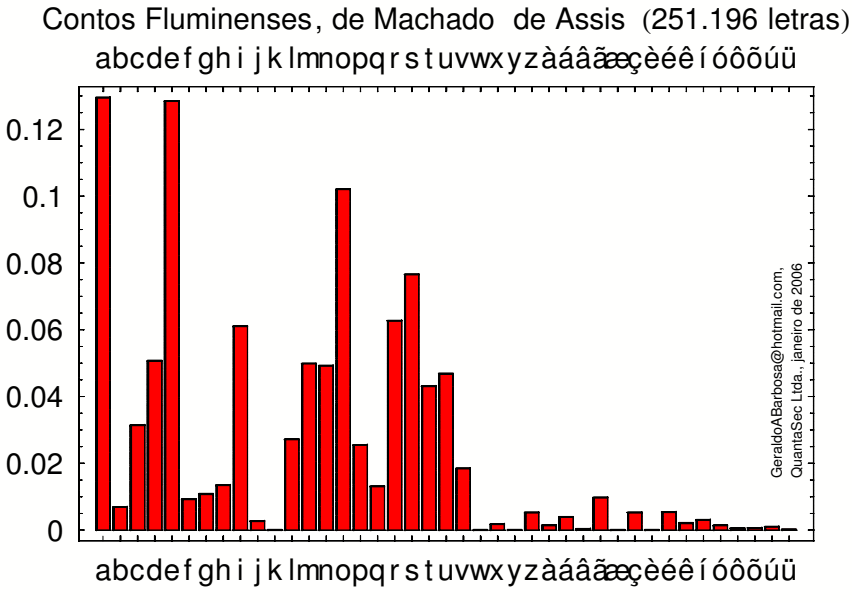


FIG. 9: Freqüência das letras nos Contos Fluminenses. A tabela à esquerda da figura apresenta numericamente as letras encontradas e sua probabilidade de ocorrência.

$$\sigma \;=\; 0,0170 \;. \tag{9}$$

Estes resultados indicam que, em amostragens bem maiores, muito possivelmente os valores a serem obtidos estarão em torno de $H \simeq 4.18 \pm 0,02$.

Bigrama de Mémorias Póstumas de Brás Cubas

O bigrama em barras do texto analisado mais detalhadamente, "Mémorias Póstumas de Brás Cubas" está mostrado na figura 10 e deve ser analisado como na figura 3. A figura 11 mostra a matriz de valores do bigrama de barras da figura 10. Deve-se observar que o alfabeto utilizado ou mostrado em cada análise pode variar, devido a utilização ou não de certas letras. De outra forma, pode-se também mostrar um alfabeto maior do que o utilizado para, eventualmente, se acomodar letras de algum alfabeto diferente (como o "ñ"). Por exemplo, a figura 10 foi construida com posições para o alfabeto de 45 letras: a1,b2,c3,d4,e5,f6,g7,h8,i9,j10,k11,l12,m13,n14,o15,p16,q17,r18,s19,t20,u21,v22,w23,x24,y25, z26,à27,á28,â29,ã30,ç31,è32,é33,ê34,ì35,í36,ñ37,ò38,ó39,ô40,õ41,ù42,ú43,û44,ü45. Neste caso, muitas letras jamais foram citadas e o bigrama mostrará o valor zero para as mesmas. A altura das barras mostra o número de ocorrências de cada bigrama. A divisão do número de ocorrências de cada bigrama, dividido pelo número total de letras com ocorrências fornece a probabilidade adjunta $P(x, y)$. A figura 11 mostra numericamente as ocorrências de cada bigrama.

CARTA DE CAMINHA

Esta seção analisa a "Carta de Caminha" nas versões contemporânea e original. Dentre os resultados buscados pergunta-se: 1) Qual é a diferença estatística na ocorrência das letras nas duas versões? 2) Qual é a entropia das duas versões e a entropia por símbolo?

Carta, versão contemporânea

A figura 12 mostra a estatística de ocorrência das letras na versão contemporânea da "Carta de Caminha" A figura 13 mostra a diferença de ocorrência de letras entre a versão contemporânea da "Carta de Caminha" e "Memórias Póstumas de Brás Cubas".

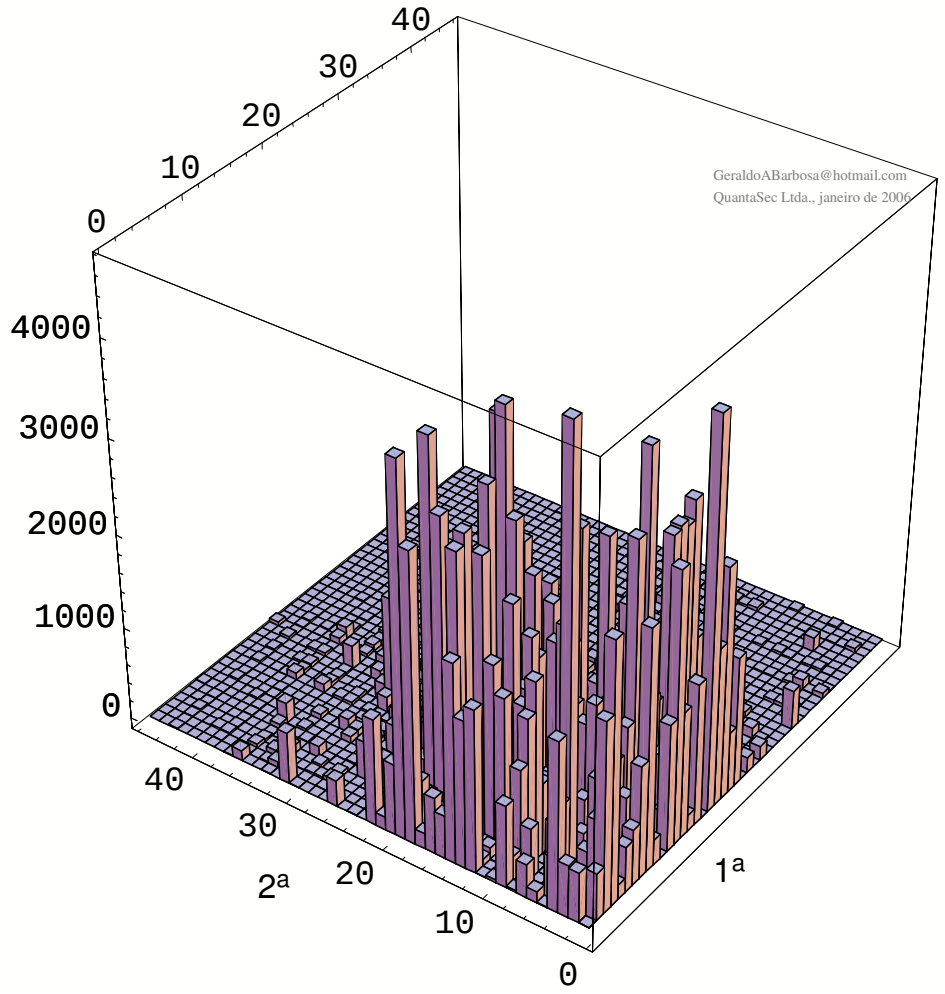


FIG. 10: Bigrama tridimensional do texto completo "Mémorias Póstumas de Brás Cubas". Por simplicidade, deve-se utilizar os números associados ao alfabeto de 45 letras utilizado.

A entropia da versão contemporânea é $H_{Cont} = 4,117$ e sua entropia por símbolo é H_{Cont}/símbolo$= 0,114$. A redundância da Carta em versão contemporânea é $r_{Cont} = 0,204$.

Carta, original

A figura 14 mostra a estatística de ocorrência das letras na "Carta de Caminha" original. A figura 15 mostra a diferença de ocorrência percentual de letras entre a "Carta de Caminha" original e "Memórias Póstumas de Brás Cubas".

A entropia da Carta original é $H_{Original} = 4,082$ e a entropia por símbolo é $H_{Original}$/símbolo$= 0,136$. A redundância da Carta original é $r_o = 0,168$.

	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	ñ	ò	ó	ô	õ	ù	ú	û	ü
a	1	540	564	1822	6	116	359	19	885	31	4	1744	1582	2131	465	606	164	3075	3927	766	137	1149	0	4	0	281	0	0	0	3	554	0	1	0	0	107	1	0	0	0	0	0	7	0	0
b	510	0	0	0	461	0	1	0	270	22	0	70	1	0	507	0	0	616	91	7	87	0	0	0	3	0	0	6	0	0	0	0	61	12	0	3	0	0	4	0	0	0	0	0	0
c	2034	2	0	0	958	0	0	543	1120	0	2	126	1	0	3074	0	0	399	0	19	477	0	0	0	0	0	0	54	21	18	4	0	47	68	0	38	0	0	20	20	2	0	6	0	0
d	2787	0	0	0	4700	0	2	5	1590	4	0	0	27	1	3190	0	2	77	0	0	212	14	0	0	2	0	0	42	1	38	0	0	88	45	0	19	0	0	10	3	1	0	18	0	0
e	152	125	752	361	44	176	506	1	1909	162	0	1586	1888	2950	34	383	65	3164	3908	538	1223	430	0	330	0	420	0	1	0	8	124	0	0	0	0	2	0	0	3	1	0	0	1	2	0
f	494	0	0	0	428	1	0	0	762	0	0	123	0	0	503	0	0	160	0	0	199	0	0	0	0	0	0	13	6	2	0	0	9	2	0	54	0	0	8	2	0	0	7	0	0
g	604	0	0	1	298	0	0	0	267	0	0	41	7	78	610	0	0	393	1	0	728	0	0	0	0	0	0	12	3	1	0	0	6	32	0	219	0	0	12	1	1	0	2	0	6
h	1210	0	0	0	865	0	0	0	75	0	0	0	0	0	972	0	0	0	0	0	175	0	0	0	0	0	0	137	0	71	0	0	3	0	0	0	0	0	0	17	1	0	0	0	0
i	2571	65	701	1036	148	105	470	0	2	66	0	454	1180	2202	897	63	45	1292	2051	1205	176	452	0	249	0	317	0	45	1	78	145	0	1	53	0	0	0	0	2	3	7	0	17	0	0
j	267	0	0	0	117	0	0	0	3	0	0	0	0	0	191	0	0	0	0	0	139	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	0	0	11	0	1	0	2	0	0
l	0	1	0	0	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	1365	0	47	32	1141	16	302	1230	976	0	0	7	162	0	1057	32	12	5	41	212	262	136	0	0	1	0	0	189	3	10	22	0	33	34	0	89	0	0	73	1	4	0	20	0	0
n	2892	310	0	0	3015	0	0	0	847	0	0	2	3	0	1543	658	8	0	0	0	470	1	3	0	0	0	0	38	1	168	0	1	36	7	0	38	0	0	32	6	0	0	3	0	0
o	1358	0	798	1359	686	231	270	1115	517	48	0	7	0	1	1264	0	67	12	690	2839	236	135	0	12	0	21	0	29	4	1188	200	0	7	3	0	26	0	0	43	29	0	0	22	0	0
p	171	386	303	488	91	122	159	1	766	48	0	829	1619	1400	7	119	10	2499	3623	254	1726	174	0	21	0	53	0	4	0	13	73	0	1	0	0	15	0	0	0	0	0	0	0	0	0
q	1516	0	2	0	1127	0	0	4	277	0	0	329	0	2	1708	0	0	974	3	7	214	0	0	0	0	0	0	49	2	9	9	0	117	6	0	256	0	0	32	27	4	0	33	0	0
r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3341	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38
s	4109	124	188	373	2739	25	422	0	1835	1	0	17	195	198	1576	46	140	523	148	678	211	80	0	0	2	1	0	164	7	25	77	0	71	89	0	34	0	0	69	2	3	0	9	0	0
t	1692	18	562	24	2774	56	24	3	704	0	0	12	278	13	1058	490	69	0	1627	1940	390	22	0	0	0	0	0	26	4	141	4	1	91	7	0	65	0	0	127	1	14	0	18	0	0
u	2471	0	8	0	2708	0	0	2	1286	0	0	2	1	0	2241	0	0	1523	0	2	786	0	0	0	0	0	0	145	37	240	0	0	152	30	0	44	0	0	31	10	4	0	7	0	0
v	646	142	161	303	2912	20	95	2	753	51	0	686	2188	554	43	120	5	575	479	613	0	142	0	49	0	58	0	9	0	0	63	0	40	23	0	72	0	0	0	0	0	0	0	0	0
w	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x	163	0	51	0	112	0	0	0	43	0	0	0	0	0	86	119	0	0	0	37	11	1	0	0	0	0	0	3	0	35	0	0	0	0	0	5	0	0	0	0	12	0	1	0	0
y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z	150	0	0	0	326	0	0	0	209	0	0	0	4	0	27	0	0	0	0	0	15	0	0	0	0	0	0	0	0	30	0	0	1	7	0	2	0	0	0	0	2	0	0	0	0
à	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
á	0	17	114	2	0	15	71	0	0	0	0	32	0	0	0	4	0	115	119	33	8	88	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
â	0	0	0	0	0	0	0	0	0	0	0	0	20	79	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ã	0	0	0	0	58	0	0	0	0	0	0	0	0	0	2498	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ç	418	0	0	0	0	0	0	0	0	0	0	0	0	0	226	0	0	0	0	0	5	0	0	0	0	0	0	3	0	528	0	0	0	0	0	0	0	0	0	0	97	0	0	0	0
è	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
é	0	25	39	33	0	2	12	0	101	0	0	7	152	2	0	14	6	131	63	20	45	2	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ê	0	3	0	0	2	0	0	0	0	0	0	1	19	216	0	0	0	0	76	2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ì	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
í	51	4	105	29	1	11	10	0	0	0	0	248	17	62	4	28	2	64	55	276	0	78	0	0	0	8	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ñ	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ò	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ó	0	7	29	21	0	11	27	0	22	0	0	25	0	0	0	43	4	84	91	18	0	4	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ô	0	0	0	9	0	1	0	0	0	0	0	0	32	44	4	1	0	5	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
õ	0	0	0	0	153	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ú	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ù	0	40	5	12	0	0	1	0	0	0	0	44	24	41	0	24	0	11	6	11	0	21	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ü	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 11: Matriz com o número de ocorrências de bigramas. Estas ocorrências são as alturas das barras da figura 10.

a	0.136973
b	0.00975453
c	0.0286484
d	0.0439686
e	0.127248
f	0.00916867
g	0.0114535
h	0.0130646
i	0.0484504
j	0.00231414
l	0.0285313
m	0.0573554
n	0.045609
o	0.105425
p	0.0223212
q	0.0172535
r	0.0612221
s	0.0898705
t	0.0447888
u	0.045814
v	0.0165505
x	0.00117171
z	0.00565352
à	0.00140606
á	0.00454039
â	0.000117171
ã	0.00934443
ç	0.00383737
é	0.00345656
ê	0.0008202
í	0.00120101
ó	0.00193333
ô	0.000322222
õ	0.000146464
ú	0.0000878786
ù	0.000175757

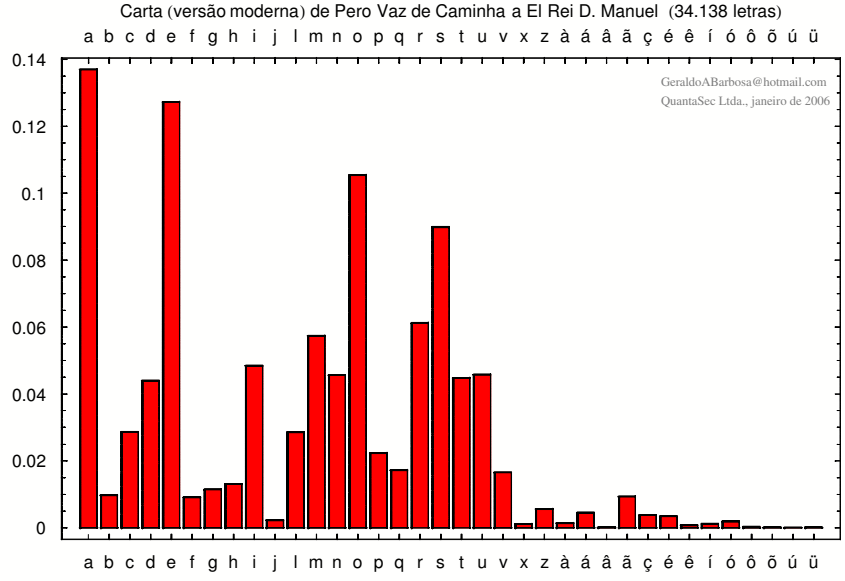
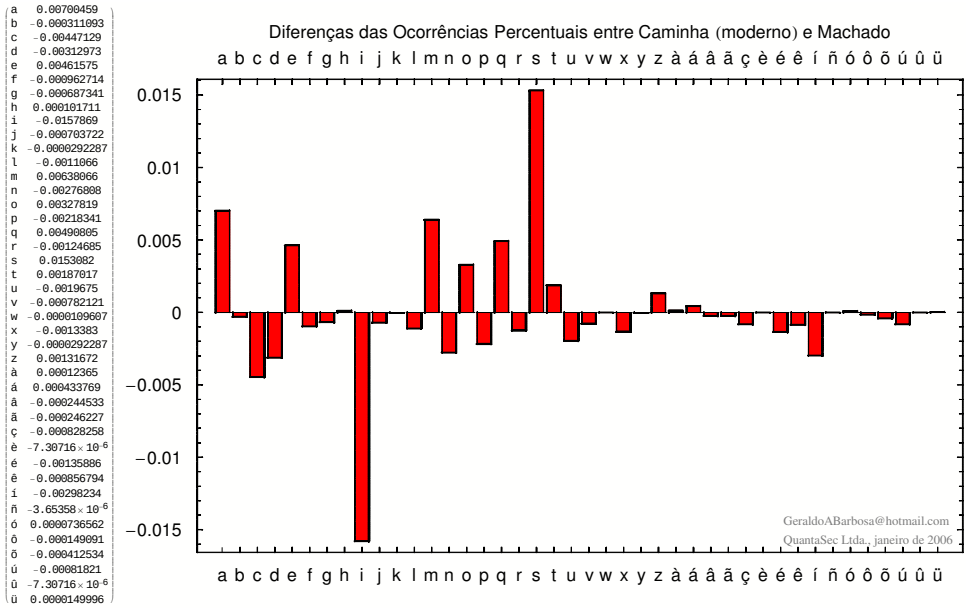
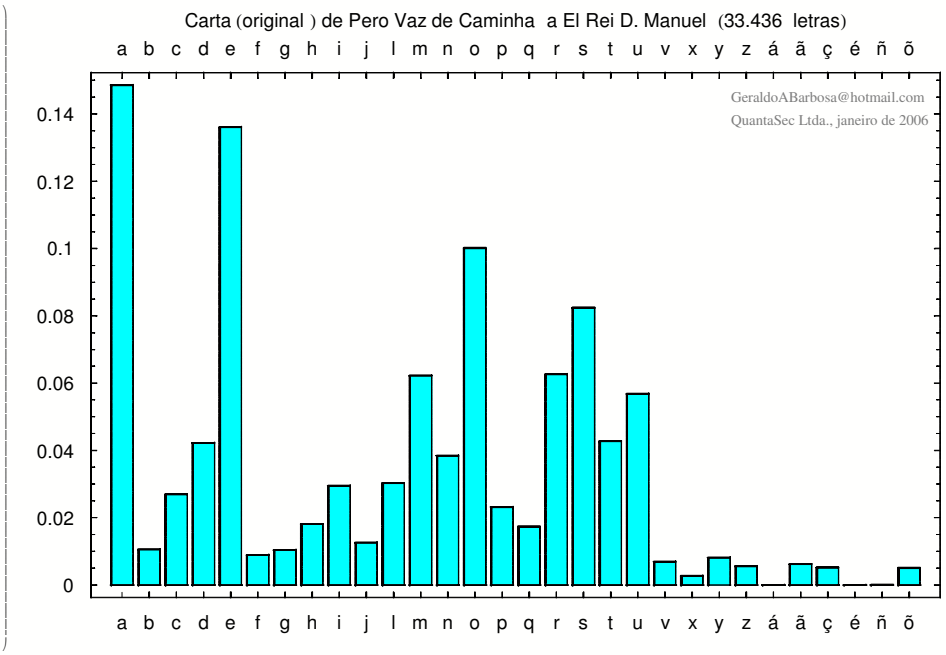


FIG. 12: Estatística de ocorrência das letras na versão contemporânea da “Carta de Caminha”. A tabela à esquerda da figura apresenta númericamente as letras encontradas e sua probabilidade de ocorrência.



a	0.00700459
b	−0.000311093
c	−0.00447129
d	−0.00312973
e	0.00461575
f	−0.000962714
g	−0.000687341
h	0.000101711
i	−0.0157869
j	−0.000703722
k	−0.0000292287
l	−0.0011066
m	0.00638066
n	−0.00276808
o	0.00327819
p	−0.00218341
q	0.00490805
r	−0.00124685
s	0.0153082
t	0.00187017
u	−0.0019675
v	−0.000782121
w	−0.0000109607
x	−0.0013383
y	−0.0000292287
z	0.00131672
à	0.00012365
á	0.000433769
â	−0.000244533
ã	−0.000246227
ç	−0.000828258
è	-7.30716×10^{-6}
é	−0.00135886
ê	−0.000856794
í	−0.00298234
ñ	-3.65358×10^{-6}
ó	0.0000736562
ô	−0.000149091
õ	−0.000412534
ú	−0.00081821
û	-7.30716×10^{-6}
ü	0.0000149996

FIG. 13: Estatística da diferença de ocorrência de letras entre a versão contemporânea da "Carta de Caminha" e "Memórias Póstumas de Brás Cubas". A tabela à esquerda da figura apresenta númericamente as letras encontradas e sua probabilidade de ocorrência.



a	0.148552
b	0.0105276
c	0.026947
d	0.0422299
e	0.136051
f	0.00891255
g	0.0104079
h	0.0180943
i	0.0294593
j	0.0126211
l	0.0302668
m	0.0622084
n	0.0384615
o	0.100162
p	0.0231487
q	0.0173765
r	0.0625972
s	0.0824261
t	0.0427085
u	0.056825
v	0.00693863
x	0.00272162
y	0.00816485
z	0.00559277
á	0.0000299079
ã	0.00619093
ç	0.00520397
é	0.0000299079
ñ	0.0000598158
õ	0.00508434

FIG. 14: Estatística de ocorrência das letras na "Carta de Caminha" original. A tabela à esquerda da figura apresenta numericamente as letras encontradas e sua probabilidade de ocorrência.

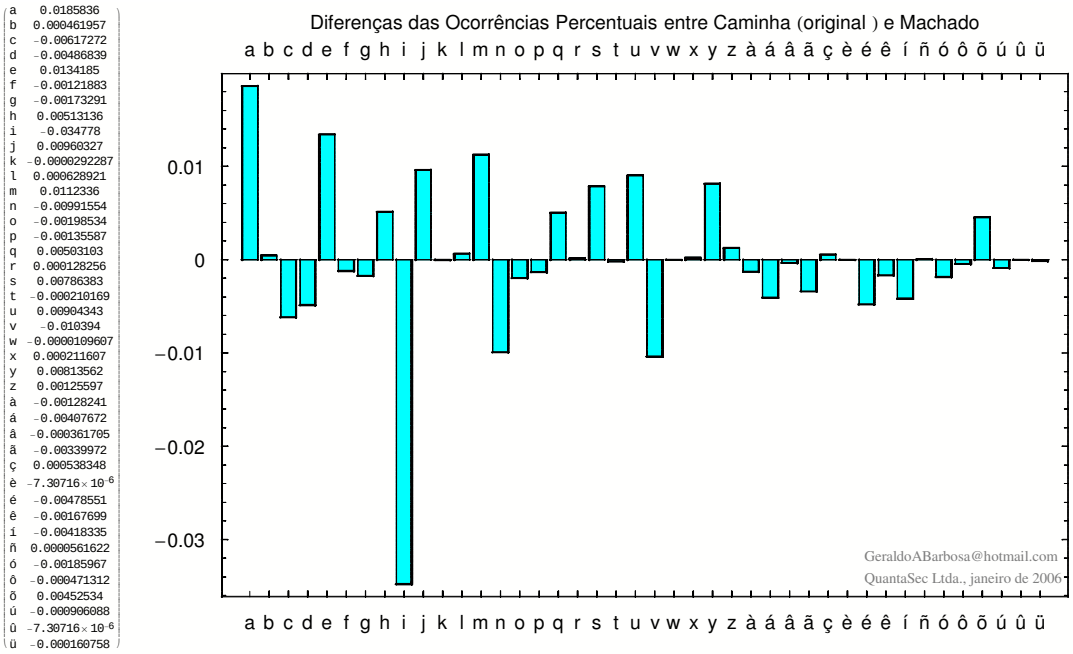


FIG. 15: Diferença de ocorrência percentual de letras entre a "Carta de Caminha" original e "Memórias Póstumas de Brás Cubas". A tabela à esquerda da figura apresenta numericamente as letras encontradas e sua probabilidade de ocorrência.

Algumas comparações

É interessante notar que as versões contemporânea e original mostram as entropias por símbolo 0.114 e 0.136, respectivamente. A redundância da versão contemporânea é maior do que a do original, $r_{Cont} = 0,204 > r_o = 0,168$, mostrando que a mesma mensagem foi transmitida por Caminha de forma mais econômica do que a versão moderna. De fato, basta observar que a mensagem original (33.436 letras) utiliza cerca de 1.000 símbolos *menos* do que a contemporânea (34.138 letras). Será que este fato deve-se mais aos redatores da forma contemporânea, em uso redundante da língua portuguesa do que uma característica da própria lingua?

A figura 16 mostra as diferenças percentuais relativas na ocorrência de letras entre a "Carta de Caminha" original e sua versão moderna. Os dados foram calculados como

$$\frac{\text{Ocorrência fracionária de letras no original-Ocorrência fracionária de letras na versão contemporânea}}{\text{Ocorrência fracionária de letras no original+Ocorrência fracionária de letras na versão contemporânea}} \times 100 \ . \qquad (10)$$

Assim, os valores 100% indicam letras existentes no original mas não na versão contemporânea e -100%, o oposto.

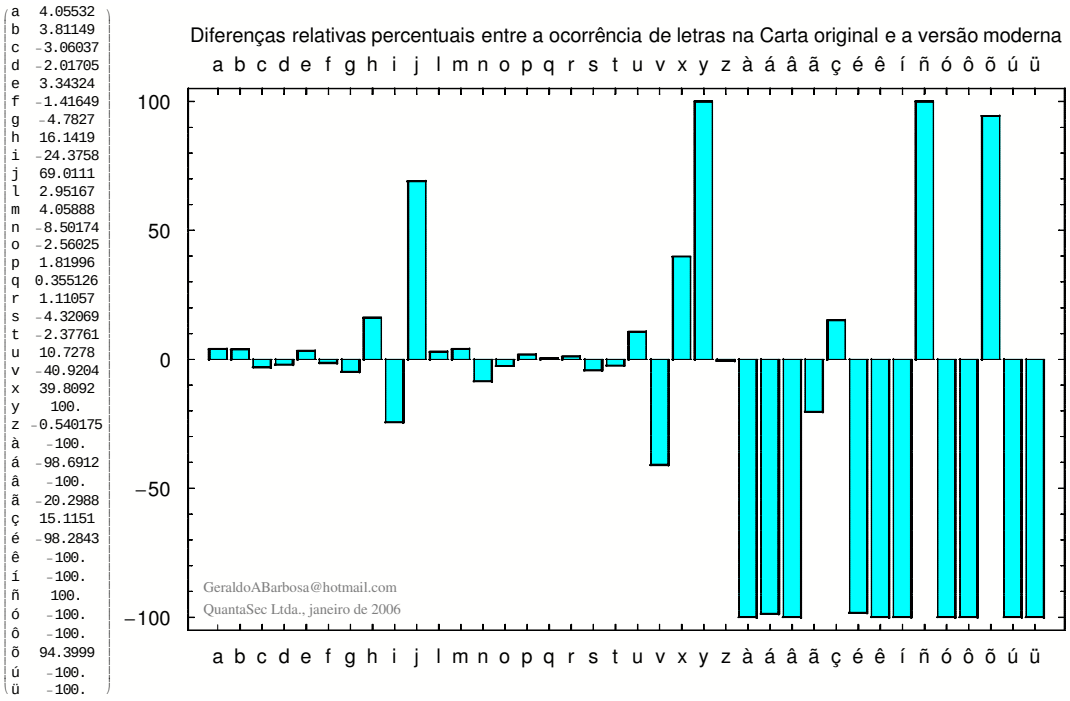


FIG. 16: Diferenças relativas percentuais na ocorrência de letras entre a "Carta de Caminha" original e sua versão contemporânea. A tabela à esquerda da figura apresenta númericamente as letras encontradas e as diferenças percentuais relativas para cada letra.

CONCLUSÕES

Neste trabalho examinou-se a estatística de ocorrência de letras na língua portuguesa utilizando-se textos de Machado de Assis acessíveis na Internet e contabilizando mais de um milhão de palavras. Para a obra "Memórias Póstumas de Brás Cubas" foi apresentado também seu bigrama. Um valor numérico representativo da *entropia* da língua portuguesa –pelo menos à época deste autor– foi calculado. A existência de apreciáveis flutuações estatísticas para amostragens em baixo número foi evidenciada assim como seu amortecimento com o aumento deste número. Foi indicado que pequenas flutuações da entropia estarão sempre presentes em obras de qualquer porte mas estas flutuações ocorrerão em torno de um valor que se pode dizer característico da língua portuguesa. Foram comparados também os resultados de Machado de Assis com os obtidos da Carta de Caminha, nas versões contemporânea e original. Mostrou-se também diferenças estatísticas entre o original e a referida versão contemporânea da Carta.

Seria muito interessante um estudo de maior número de textos visando corroborar ou

não os resultados aqui obtidos e se estabelecer de forma definitiva alguns parametros significativos da língua portuguesa contemporânea e mesmo a evolução da língua. A obtenção de entropia em melhores aproximações poderia ser também um dos objetivos nesta análise maior, para fornecer um estudo bem mais completo.

O advento da computação possibilitando o tratamento de grandes volumes de dados indica que mudanças introduzidas na língua por propostas conscientes deveriam ser acompanhadas por estudos estatísticos que apoiassem estas propostas.

Mais do que os resultados aqui apresentados pretende-se estimular os jovens interessados na língua portuguesa a se utilizarem de ferramentas estatísticas que se tornam indissociáveis do mundo moderno e que permitem análises impraticáveis de serem feitas manualmente.

*O autor é Ph.D. pela University of Southern California (1974) e foi Professor Titular da Universidade Federal de Minas Gerais, Brasil. Atualmente, é Professor no Center for Photonic Communication and Computing, ECE Department, Northwestern University, Evanston, IL 60208-3118, US

---

* Electronic address: Email:GeraldoABarbosa@hotmail.com,g-barbosa@northwestern.edu

[1] D. Kahn, *The Code-Breakers, The Story of Secret Writing* (Scribner, New York 1996). S. Singh, *The Code Book, The Science of Secrecy from Ancient Egypt to Quantum Cryptography* (Anchor Books, New York 1999).

[2] J. Bamford, *Body of Secrets, Anatomy of the Ultra-Secret National Security Agency* (Anchor Books, New York 2002).

[3] D. J. C. MacKay, "*Information Theory, Inference, and Learning Algorithms*" (Cambridge 2003).

[4] C. E. Shannon, The Bell System Technical Journal, Vol. **27**, pp. 379423, 623656, July, October, 1948.