Maria Madalena de Freitas Malva

Quem Foi Que?
— Um Desafio à Estatística:
Questões de Autoria em "Novas Cartas Portuguesas"

Dissertação submetida para satisfação parcial
dos requisitos do programa de Mestrado em
Probabilidades e Estatística.

Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional
Dezembro 1998

Ao Prof. Doutor Dinis Pestana desejo expressar o meu reconhecimento por me ter proposto o tema deste trabalho, pela sua orientação, pela confiança, pelo apoio e pela permanente disponibilidade durante a sua realização.

À Profª. Doutora Emília Athayde desejo agradecer as sugestões sugeridas pois permitiram uma melhoria formal deste trabalho.

À Doutora Maria Isabel Barreno quero agradecer o interesse demostrado por este trabalho.

Ao Departamento de Informática e ao Departamento de Matemática da Escola Superior de Tecnologia de Viseu o meu muito obrigado por me terem disponibilizado os meios informáticos necessários para a realização deste trabalho.

A todos os meus amigos e familiares desejo agradecer o interesse e a paciência demonstrados, em particular aos meus pais e à Maria de Céu por terem sido quem mais de perto me acompanhou "nesta aventura". A todos a minha mais profunda gratidão.

Índice

Nota Introdutória

Embora os problemas de disputa de autoria sejam comuns na história, literatura e política e o seu estudo se revista de grande interesse, não têm despertado muito a atenção da comunidade estatística em geral, ou pelo menos, não se tem verificado um grande desenvolvimento nesta área. O facto de até há bem pouco tempo não existirem computadores e programas que permitissem uma análise rápida e eficiente dos textos, nomeadamente no que diz respeito à contagem de frequências de palavras, fez com que os problemas de autoria se tornassem pouco atractivos para os estatísticos. De um modo geral, estes problemas são tratados como problemas de discriminação ou classificação, onde o que se pretende é fazer a atribuição de uma categoria (autor) a um objecto ou indivíduo (texto) cuja verdadeira categoria é desconhecida.

Alguns estudos estatísticos sobre problemas de autoria foram efectuadas ao longo destes últimos anos, sendo o mais famoso o dos “Papeis Federais” efectuado por Mosteller e Wallace, [8]. Neste estudo os autores usam várias técnicas e abordagens estatísticas, inclusive a Bayesiana, dando assim o seu contributo para a resolução do problema da atribuição da autoria de doze artigos, disputada por Madison e Hamilton. James Madison e Alexander Hamilton juntamente com Jonh Jay escreveram setenta e sete artigos para persuadir os cidadãos do estado de Nova York a ratificar a constituição, artigos esses que foram publicados em vários jornais sob o pseudónimo de “Publius”. Dos setenta e sete artigos publicados sabe-se que cinco foram escritos por Jay, que Hamilton escreveu quarenta e três e Madison catorze; três foram escritos por Hamilton e Madison em parceria, sendo os restantes disputados por Hamilton e por Madison. Foi sobre estes artigos que Mosteller e Wallace se debruçaram tentando chegar a uma conclusão sobre a sua autoria.

Outros problemas famosos nesta área relacionam-se com o estudo do Antigo e do Novo Testamento. Uma das questões que se coloca neste campo tem a ver com as epístolas de S. Paulo, já que se coloca a questão de se saber se estas foram todas

escritas pelo mesmo autor. Outra questão polémica respeita o Antigo Testamento e tem por base o livro de Ìsaias; existem estudiosos que afirmam que este livro foi escrito por dois autores: um terá escrito os capítulos 1-39 e o outro os capítulos 40-66. Este tipo de problemas não se encontra apenas na história e na literatura mas também na vida corrente. De facto, por exemplo, a falsificação de documentos é usual e, por vezes, os métodos utilizados em investigação criminal não são suficientes para resolver o problema, pelo que o estatístico poderá também, neste campo, ter um importante contributo a dar.

A consideração de um tipo irónico de falsificação, o "pastiche", mostra bem a que ponto a intuição pode guiar um bom autor a imitar outro(s), por vezes enganando mesmo os especialistas. Basta referir Chatterton e a sua falsificação de um bardo medieval, ou o fabuloso "Pastiches et Mélanges", de Proust.

Seria aliás curioso uma abordagem estatística ao plágio, a imitação que se esconde (agora: no tempo de Camões, plagiar Petrarca ou os clássicos nada tinha de embaraçoso), e ao "pasticha", a imitação caricatural que se assume.

Os problemas da atribuição de autoria são pois abundantes e as suas soluções revestem-se de muito interesse, pelo que seria de esperar um maior interesse por este tipo de problemas por parte da comunidade estatística.

Várias questões se colocam quando nos deparamos com este tipo de problemas; de entre todas, talvez a mais importante seja a da definição das variáveis a utilizar, pois, sem a sua definição e quantificação, nenhum estudo pode avançar. Na bibliografia consultada sobre este assunto, um denominador comum reside no facto de as variáveis se encontrarem todas bem definidas e parecer extremamente natural e lógico utilizar tais variáveis — tudo parece fazer sentido. O caminho para chegar a tais variáveis é, de um modo geral, omitido ou então apresentado como sendo simples de percorrer. Os problemas começaram a surgir quando, perante um problema de atribuição de autoria nunca antes abordado, se tem que definir as variáveis a utilizar; o que fora apresentado como um caminho simples torna-se, de súbito, um caminho tortuoso e/ou um beco sem saída, pois na maior parte das vezes uma variável que parecia ser promissora como identificadora do autor revela-se um

verdadeiro fiasco. Uma variável pode ser excelente numa situação e não ter qualquer utilidade noutra, pois a definição das variáveis a utilizar depende dos autores em estudo e, por vezes, até do contexto histórico-social em que estes estão inseridos. Outro aspecto importante a ter em conta na definição e quantificação das variáveis é a grande subjectividade que a elas pode estar associada. Por exemplo, se se pretender estudar o modo como as orações se encontram ligadas numa frase, verifica-se que estas podem estar ligadas por coordenação e/ou por subordinação, mas tal classificação não é tão simples de fazer como à primeira vista pode parecer; casos há de frases em que nem os especialistas conseguem estabelecer o modo de ligação entre as orações.

Assim, perante um problema de atribuição de autoria, várias questões se devem colocar aquando da definição das variáveis. Por exemplo, será a frequência das palavras uma variável a utilizar? Se sim, será que todas as palavras têm interesse, ou só uma parte das palavras utilizadas pelo autor são verdadeiramente importantes, e, neste caso, quais as palavras que se devem estudar? Serão o tamanho de frase e o tamanho de parágrafo variáveis a considerar? Será a o tipo pontuação utilizada uma variável distintiva do autor? E que outras variáveis poderão ser importantes para definir e identificar o estilo de um autor?

O objectivo deste trabalho é mostrar que em problemas de atribuição de autoria não é fácil escolher e quantificar as variáveis a utilizar. Para tal vai-se considerar um conjunto de variáveis, ou mais precisamente um conjunto de candidatas a variáveis, tais como: frequência de determinadas palavras, o comprimento de frase e o comprimento de parágrafo, tipo e frequência de pontuação utilizada, o modo como as orações estão ligadas nas frases, etc.. Seguidamente, vão-se aplicar algumas técnicas de análise exploratória de dados e verificar até onde estas nos conduzem e que conclusões é lícito retirar. O objectivo último é chegar a um conjunto de variáveis que identifique o autor, para se poder avançar no sentido da atribuição de autoria de textos “desconhecidos”. Neste trabalho estuda-se, entre outras, as variáveis atrás enunciadas, tendo como objectivo final, se possível, fazer uma atribuição de autoria.

Como corpus de estudo decidiu-se utilizar o livro "Novas Cartas Portuguesas" escrito de parceria por Maria Isabel Barreno, Maria Teresa Horta e Maria Velho da Costa, [2]. Este livro é constituído por um conjunto de cartas e poemas que ecoam a história de Mariana Alcoforado, freira num convento de Beja e da qual se diz que teve um caso amoroso com um cavaleiro francês de nome Chamilly. O livro foi publicado em 1972 e gerou, na época, grande polémica; esteve inclusivé proibido pela DGS - Direcção Geral de Segurança, devido à linguagem nele utilizada ser considerada, à época, como imprópria e lesiva da moral. O caso ficou conhecido na literatura e na história como o "Caso das Três Marias". Nenhuma das autoras assinou qualquer das cartas ou poemas, nem assumiu a sua autoria, e, tanto quanto sabemos, nunca se fez qualquer estudo para esclarecer a autoria dos textos.

1. Alguns conceitos básicos de análise exploratória de dados

As técnicas de análise exploratória de dados permitem de forma fácil e rápida conceber a estrutura dos dados. Antes de produzir estatísticas ou formular hipóteses à custa dos dados, convém examinar detalhadamente os mesmos. Esta necessidade de examinar pormenorizadamente os dados é sentida de modo particular quando os dados em estudo provêm de estudos de autoria.

Quando se pensa em sintetizar uma colecção (amostra) de dados constituída por n observações x_1, x_2, ..., x_n as estatísticas-resumo clássicas mais usuais são a média aritmética $\overline{x}$, dada por

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

e a variância amostral,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 .$$

A média aritmética e a variância amostral são duas ferramentas poderosíssimas que estão na base de alguns dos resultados mais fortes da teoria das probabilidades e estatística. Pense-se, por exemplo, na desigualdade de Tchebychev: "seja X uma variável aleatória real positiva e seja g uma função crescente de $\Re^+$ em $\Re^+$ tal que E(g(X)) existe; então

$$\text{para qualquer } \alpha > 0,\ P(X \geq \alpha) \leq \frac{E(g(X))}{g(\alpha)}\ ".$$

Ou ainda, na desigualdade de Biénayme-Tchebychev cujo enunciado diz: "se X é uma variável aleatória real admitindo momento de segunda ordem finito, então

$$\text{para qualquer } \alpha > 0,\ P(|X-E(X)|\geq\alpha) \leq \frac{Var(X)}{\alpha^2}\text{".}$$

Os resultados anteriormente apresentados, são suficientemente fortes e importantes, para demonstrarem de uma forma clara a importância da média aritmética e da variância amostral.

Contudo, quando perante uma colecção de dados sobre a qual se pretende ter uma ideia sobre a sua localização e sobre a sua dispersão, nem sempre a média e a variância são as estatísticas mais adequadas para descrever estas características. Casos há, em que a média e a variância dão uma ideia completamente errada sobre a localização e sobre a dispersão da colecção. Quando a colecção em estudo tem um ou mais valores que fogem ao "padrão geral da colecção", as duas estatísticas-resumo, referidas anteriormente, são fortemente influenciadas por estes valores, dando uma ideia, por vezes distorcida, sobre a localização e sobre a dispersão da amostra. Tal deve-se ao facto de, quer a média quer a variância, serem estatísticas-resumo pouco resistentes, i.e., uma mudança arbitrária numa pequena parte da colecção pode ter um efeito adverso substancial em qualquer uma destas estatísticas.

Assim, por vezes, é mais vantajoso utilizar estatísticas-resumo mais simples baseadas na ordenação e contagem dos dados, mas, que de um modo geral, sejam mais resistentes, i.e., uma mudança arbitrária numa pequena parte da colecção tem só um pequeno efeito na estatística-resumo. A mediana, i.e., a estatística que está a meio caminho da amostra ordenada, e a dispersão-quartal definida por:

Dispersão-Quartal = (quarto superior) - (quarto inferior),

que fornece o comprimento da metade central da colecção de dados, são exemplo de duas estatísticas-resumo resistentes para a localização e para a dispersão, respectivamente.

Apesar da fraca resistência da média aritmética e da variância, neste trabalho não se resistiu à tentação de, perante colecções de dados, se calcular a sua média e o seu desvio-padrão ($S=+\sqrt{S^2}$), para assim se ficar com uma ideia geral sobre a localização e sobre a dispersão das várias colecções, e deste modo ser possível estabelecer comparações entre estas. No entanto, devido à pouca resistência destas duas estatísticas nem sempre se ficou pelo seu cálculo. Sempre que a colecção em causa apresentava possíveis outlieres, i.e., observações estranhas, foi calculado o valor da mediana e da dispersão-quartal, bem como outras estatísticas-resumo, como o valor do máximo e/ou do mínimo.

Quando se fala em análise exploratória de dados pensa-se logo em representações gráficas, pois estas fornecem uma impressão visual de vários aspectos importantes da distribuição empírica de uma colecção de dados. A caixa-com-bigodes é uma das representações gráficas mais utilizadas na análise exploratória de dados. Esta representação é construída à custa de cinco estatísticas-resumo: a mediana, o quarto-inferior, o quarto-superior, o extremo inferior e o extremo superior e revela grande parte da estrutura dos dados. De uma caixa-com-bigodes podem extrair-se as seguintes características de uma colecção:

Localização

Dispersão

(As)simetria

Comprimento de Caudas

Outliers.

Esta representação visual compacta é especialmente útil para comparar diferentes colecções de dados. Ao dispor-se em paralelo caixas-com-bigodes para cada uma das colecções, podem-se comparar as localizações e as escalas das diferentes colecções, e talvez mesmo a assimetria e o peso das caudas. Pode-se assim concluir, desta simples comparação, que os dados das diferentes colecções não se ajustam bem à mesma escala. Em particular, as colecções localizadas longe da origem podem eventualmente ser muito mais dispersas do que as colecções localizadas perto da origem. Portanto se as colecções forem marcadas numa escala comum, será mais difícil detectar peculiaridades dos dados numa vizinhança da origem.

Uma transformação permite frequentemente ultrapassar esta dificuldade, tornando mais homogénea a variabilidade das colecções. Quando a comparação de várias colecções mostra a existência sistemática de uma relação entre a dispersão e o nível, tenta-se frequentemente encontrar uma re-expressão, ou transformação, dos dados originais que reduza ou elimine esta dependência. Se se conseguir encontrar tal transformação, os dados re-expressos serão mais convenientes quer para exploração visual quer para as técnicas analíticas mais usuais de comparação de colecções.

Transformação-potência

Define-se transformação potência como uma transformação da forma:

$$T_p(x)=\begin{cases} ax^p+b & \text{se } p\neq 0 \\ c\log x+d & \text{se } p=0 \end{cases},$$

com a, b, c, d e p reais.

Prova-se que as transformações potência $T_p(x)$ gozam das seguintes propriedades:

- Preservam a ordem dos dados, i.e., são funções estritamente crescentes.
- Preservam as estatísticas-resumo de uma colecção de dados a menos de diferenças negligíveis resultantes de interpolações.
- São contínuas.
- São funções regulares, com derivadas de todas as ordens.
- São funções elementares.

As constantes a, b, c e d são, de um modo geral, arbitrárias; apenas a primeira propriedade referida, está dependente da escolha das constantes aditivas ou multiplicativas; e é válida, no caso transformação-potência, se se multiplicarem as potências de expoente positivo por constantes positivas, e potências de expoente negativo por constantes negativas, ou seja, $a>0$ para $p>0$ e $a<0$ para $p<0$. A escolha de p é feita por conveniência dos dados.

O contexto em que se utiliza a transformação é um guia para a determinação das constantes a, b, c e d. Apresentam-se três situações vulgares:

1. Quando se pretende re-expressar uma colecção de dados de forma particularmente simples; neste caso

$$T_p(x)=\begin{cases} x^p & \text{se } p>0 \\ \log x & \text{se } p=0 \\ -x^p & \text{se } p<0 \end{cases} .$$

2. Quando se pretende comparar transformações entre si e examinar propriedades matemáticas e geométricas, selecciona-se

$$T^*_p(x)=\begin{cases} \dfrac{x^p-1}{p} & \text{se } p\neq 0 \\ \log x & \text{se } p=0 \end{cases} .$$

3. Quando se deseja re-expressar uma colecção de valores de tal forma que os dados transformados se assemelhem aos dados originais em localização e dispersão, as constantes são seleccionadas recorrendo ao conceito de concordância.

Uma vez fixado p, qualquer escolha das constantes a e b (ou c e d, com p=0) é transformação linear de qualquer outra escolha de constantes.

É pois evidente que a escolha de constantes é determinada por conveniência e simplicidade de análise e interpretação, e não por necessidade ou alteração essencial do comportamento dos dados — de facto, qualquer transformação linear altera apenas a posição da origem e a escala, uniformemente.

Neste trabalho foram apenas utilizadas transformações do tipo das indicadas no ponto 1.

O gráfico dispersão-versus-nível é um óptimo guia da transformação a efectuar. As transformações-potência não são a única maneira de re-exprimir os dados, mas constituem uma família conveniente de transformações úteis numa grande diversidade de situações.

Construção do gráfico dispersão-versus-nível

O que se pretende, do ponto de vista heurístico, é remover a relação existente entre a dispersão e o nível. Para melhor se perceber essa relação parece razoável representar uma medida de dispersão versus medida de nível.

Admita-se, por exemplo, que a dispersão-quartal (d_F) é proporcional a uma potência da mediana (M), i.e.:

$$d_F = cM^b \text{ , com } c \neq 0 \text{ constante}$$

ou

$$\log(d_F) = \log(c) + b\log(M)$$

ou, considerando k=log (c),

$$\log(d_F) = k + b\log(M). \quad (1)$$

Tem-se, assim, uma relação linear entre os logaritmos das dispersões-quartais e os logaritmos das medianas. O gráfico de dispersão-versus-nível tem por base a asserção anterior.

O objectivo é utilizar a gráfico dispersão-versus-nível para determinar o valor de b na equação (1). A transformação $z=x^{1-b}$ dos dados x fornece os valores re-expressos, z, cujas dispersões-quartais não dependem, pelo menos aproximadamente, dos níveis.

Assim, se b for a inclinação da recta ajustada aos pontos do gráfico dispersão-versus-nível, então p=1-b é o valor aproximado do expoente da transformação-potência de x que estabiliza a dispersão. Quando p=0, utiliza-se o logaritmo dos dados.

Das características que se podem extrair de uma caixa-com-bigodes, a assimetria foi uma das características que mais atenção mereceu neste estudo. Para algumas das variáveis que se apresentam neste estudo, calculou-se o seu coeficiente de assimetria; tal cálculo foi motivado pela constatação do facto de que a diferença entre o valor da média e o valor da mediana era substancial. A assimetria de uma colecção pode estudar-se considerando a posição relativa da média, da mediana e da moda (observação mais frequente). Em colecções simétricas as três medidas de localização coincidem, e nas colecções assimétricas, a moda, a mediana e a média seguem-se por esta ordem para o lado mais longo ou menos abrupto.

Com efeito, a média desloca-se para esse lado à medida que a assimetria se acentua, porque para lá se desloca também o centro de gravidade. A mediana, como divide a área em duas iguais, para compensar a redução da área no lado abrupto, afasta-se também da moda, mas menos do que a média.

Em resumo:

1. Em colecções simétricas

média = mediana = moda.

2. Em colecções enviesadas à esquerda ou assimétricas positivas

Média > mediana > moda.

3. Em colecções enviesadas à direita ou enviesadas negativas

Média < mediana < moda.

A assimetria é tanto maior quanto mais afastadas tiverem média, mediana e moda.

Existem vários coeficientes que medem o valor da assimetria; o mais conhecido talvez seja o coeficiente de assimetria de Pearson, dado por:

$$g = \frac{\overline{X} - \text{mod}}{S}.$$

Como neste trabalho se utilizou o package “Statistica” para Windows — versão 4.5, os coeficientes de assimetria apresentados, foram calculados através da fórmula implementada no referido software. A fórmula utilizada no “Statistica” é:

$$\text{coeficiente de assimetria} = \frac{nM^3}{(n-1)(n-2)\sigma^3},$$

onde

n — número de casos válidos

$M^3 = \sum_{i=1}^{n} (x_i - \overline{X})^3$ — momento centrado de ordem três

σ — desvio padrão.

As noções atrás apresentadas, são as necessárias para que o leitor desta monografia tenha uma perfeita compreensão do trabalho efectuada e das conclusões a que se chegou.

2. Recolha de dados

Num trabalho que tem como objectivo último a atribuição da autoria a textos cujos autores são desconhecidos, mas sobre os quais recaem algumas suspeitas, é natural e lógico pensar-se em utilizar variáveis que caracterizem o estilo literário dos possíveis autores e que, directa ou indirectamente, tenham algo a ver com noções gramaticais, uma vez que estas são essenciais para definir o estilo de um escritor.

Se o autor A utiliza, nos seus escritos, determinada palavra com mais frequência do que o autor B nos seus; perante um texto, cuja autoria se discuta entre os autores A e B, se este tiver uma frequência elevada da palavra em causa, é natural dizer-se que provavelmente o autor do texto é A; caso contrário, dir-se-á que provavelmente, é B. Ou seja, a frequência das palavras é importante para este tipo de problemas, mas, vários problemas se colocam ao uso da frequência das palavras. Nomeadamente, a frequência das palavras deve ser obtida em que tipo de textos? Grandes ou pequenos? Antigos ou actuais? Será que para o autor A se devem utilizar textos grandes e actuais e para o autor B textos pequenos mas antigos? Ou é indiferente, desde que se obtenha a frequência das palavras? E, logo à cabeça — que palavras (pois o contexto decerto determina o vocabulário usado)?

Nesta secção, mostra-se como é que, neste trabalho, se contornaram as questões anteriores e outras, que foram surgindo ao longo do processo de recolha de dados.

Como já foi referido utilizou-se como corpus de estudo o livro "Novas Cartas Portuguesas", escrito de parceria por Maria Isabel Barreno, Maria Teresa Horta e Maria Velho da Costa. Para, de algum modo, se poder definir e identificar estatisticamente o estilo literário das autoras, foi necessário conhecer outros dos seus textos. Para isso utilizaram-se as obras "Os Outros Legítimos Superiores", publicado em 1970, da Maria Isabel Barreno,[1], "Ambas as Mãos Sobre o Corpo", também publicado em 1970, da Maria Teresa Horta, [7] e "Maina Mendes" da Maria Velho da Costa, [4], publicado em 1969. A escolha destas obras teve por base o facto de terem sido publicadas (e, portanto, escritas), antes da edição das "Novas Cartas

Portuguesas", que ocorreu em 1972, evitando-se assim a possibilidade do estilo de uma das autoras ter sido influenciado pelo estilo das restantes, depois do trabalho em parceria. Outro aspecto importante que se teve em consideração foi o facto de, embora, terem sido editados antes das Novas Cartas, a edição dos três livros não ter ocorrido muito tempo antes da edição das cartas, evitando-se deste modo, possíveis evoluções nos seus estilos que os tornassem dificilmente cotejáveis com anteriores padrões. Saliente-se ainda que "Ambas as Mãos Sobre o Corpo" é, tanto quanto julgamos saber, o único livro de prosa de autoria de Maria Teresa Horta, que foi publicado, ficando a escolha, no caso desta autora, limitada à obra atrás referida; esta autora é porém bastante conhecida pela sua extensa obra poética.

Sublinhe-se ainda que os livros referidos anteriormente são citados nas Novas Cartas. Como segundo título das "Novas Cartas Portuguesas" pode ler-se:

> "(ou de como Maina Mendes pôs ambas as mãos sobre o corpo e deu um pontapé no cú dos outros legítimos superiores)".

Tal facto veio reforçar, ainda mais, a ideia de que estas obras deviam ser as utilizadas no presente estudo.

2.1. Escolha dos textos de autoria conhecida

Depois de escolhidas as obras a estudar, a questão que se colocou foi a de se saber quantas e quais as partes dos livros a utilizar, uma vez que utilizar a totalidade dos livros estava fora de questão, entre outras, por razões logísticas.

Mosteller e Wallace, nos papeis federais, aconselham os indivíduos que tenham em mãos problemas de atribuição de autoria a utilizarem blocos de texto com sensivelmente o mesmo tamanho, pois tal vai facilitar a análise estatística, referindo ainda que blocos de texto com mil palavras são pequenos e com mais de três mil, um desperdício; i.e., o "tamanho ideal" dos blocos de texto a analisar situa-se assim entre as mil e as três mil palavras.

No caso em análise, existia à partida um condicionalismo, resultante da "falta de material" para Maria Teresa Horta; o livro escolhido para esta autora é constituído por trinta e nove textos, alguns dos quais "muito pequenos", colocando-se deste modo o problema da "falta de palavras". Verificou-se que a utilização de blocos com mais de mil e quinhentas palavras tinha como consequência a construção de um número reduzido de blocos para esta autora; resolveu-se então considerar blocos de mil e quinhentas palavras. Pelo que foi possível deste modo construir onze blocos de texto para a Maria Teresa, tendo-se, ainda assim, utilizado todo o livro.

Para as outras duas autoras, o problema de existir material disponível não se colocou, pois as obras que foram escolhidas são "mais extensas". Assim, para estas duas autoras consideraram-se doze blocos de mil e quinhentas palavras cada; a utilização de doze blocos ficou a dever-se ao facto de existir mais material disponível, pelo que se resolveu utilizar mais um bloco.

Decidido o número e o tamanho dos blocos a utilizar, o problema que se colocou a seguir foi o de decidir como construir tais blocos. Para a Isabel Barreno e para a Maria Velho recorreu-se ao gerador de números aleatórios do Excel, para determinar os números que iriam corresponder às páginas onde começariam os blocos de texto a estudar. Isto é, com o Excel gerava-se um número aleatório que iria corresponder ao número da página (se existisse) onde um bloco de mil e quinhentas palavras iria ter início; sempre que o número sorteado correspondesse a uma página já englobada nalgum bloco, esse número era ignorado. A tabela abaixo apresenta as páginas que constituem os doze blocos de texto considerados para estas duas autoras.

	Isabel Barreno	**Maria Velho**
Bloco 1	28-33	23-27
Bloco 2	35-40	62-69
Bloco 3	43-49	74-81
Bloco 4	106-111	89-97
Bloco 5	112-118	109-113
Bloco 6	119-124	119-125
Bloco 7	163-168	133-137
Bloco 8	169-175	165-169
Bloco 9	176-181	170-176
Bloco 10	182-187	177-181

Bloco 11	189-194	188-192
Bloco 12	196-202	220-224

Tabela 1 : Páginas que constituem os blocos estudados

No caso do livro da Teresa Horta, que foi utilizado na totalidade, o que se fez foi atribuir um número (de um a trinta e nove) a cada texto, e depois escolher aleatoriamente quais iriam constituir um bloco de mil e quinhentas palavras. Os onze blocos construídos foram:

	Textos
Bloco 1	O Andar, A Tarde, A Imobilidade e A Memória
Bloco 2	Party, O Desalento, A Mancha e A Sede
Bloco 3	Movimentos, A Infância e O Tempo
Bloco 4	O Torpor, O Ócio e O Silêncio
Bloco 5	Movimentos, O Jantar, O Vinho e O Crepúsculo
Bloco 6	O Medo, O Desespero, O Vazio e O Diário
Bloco 7	A Noite, A Cidade e O Sono
Bloco 8	A Morte, A Piscina, O Odor e O Exílio
Bloco 9	A Tarde, Os Outros, O Silêncio e A Resolução
Bloco 10	Movimentos, A Mãe e A Fuga
Bloco 11	O Banho, A Irmã e O Encontro

Tabela 2: Textos que constituem os blocos estudados

Note-se que, no caso da Teresa Horta, quando o número de palavras num bloco excedia as mil e quinhentas, as excedentes eram retiradas e utilizadas para completar blocos onde, porventura, o número de palavras fosse inferior a mil e quinhentas.

No total consideraram-se $12 \times 1500 = 18000$ palavras para a Isabel Barreno e para a Maria Velho, e $11 \times 1500 = 16500$ palavras para a Teresa Horta, números que pareceram ser perfeitamente razoáveis para efectuar o estudo proposto, uma vez que, só para os textos conhecidos foram estudadas um total de trinta e quatro mil e quinhentas palavras.

Antes de terminar este parágrafo saliente-se o facto de terem sido eliminados dos textos em estudo citações e excertos escritos em Inglês e/ou Francês.

2.2. Escolha dos textos de autoria desconhecida

Como já foi referido, o livro "Novas Cartas Portuguesas" é constituído por um conjunto de textos em prosa e poesia. Por questões que se prendem com o tempo disponível para fazer este trabalho, e não só, não se estudaram todos os textos que compõem o livro em questão.

A primeira decisão tomada na escolha dos textos "desconhecidos" a estudar, foi a de eliminar deste estudo os poemas. Tal decisão teve por base dois factores. O primeiro teve a ver com o facto de a Maria Teresa Horta ser a única autora, de entre as três em estudo, que tem poesia publicada (pelo menos do nosso conhecimento), sendo sempre referida na área da literatura como poetisa. A constatação deste facto induziu-nos a pensar (porventura erradamente) que os poemas que aparecem ao longo das Novas Cartas são de sua autoria. O segundo factor teve a ver com a maneira de escrever prosa e poesia; de facto, o estilo que está subjacente à escrita de prosa é, de modo geral, diferente do que está subjacente à escrita de poesia, pelo que não se pode comparar directamente a prosa e a poesia. Mesmo não colocando a hipótese de que Maria Teresa foi a autora dos poemas, não os deveríamos utilizar neste estudo, uma vez que se desconhece a existência de poemas da Maria Isabel e da Maria Velho para se poderem estabelecer comparações entre os "conhecidos" e os "desconhecidos". Comparar dados obtidos em textos em prosa com dados obtidos em poesia não é adequado, pois, de um modo geral, os dados obtidos em textos poéticos estão enviesados, devido aos fortes condicionalismos subjacentes à escrita de poesia. Mosteller e Wallace, nos estudo efectuado sobre os papeis federais, chamam a atenção para este facto e aconselham mesmo a não se comparar prosa com poesia.

Excluídos os poemas houve que escolher os textos em prosa a analisar. Neste ponto, outra hipótese foi colocada, por ser perfeitamente lógica. Partiu-se do princípio de que as três autoras se apresentaram e se despediram "desta aventura", e como tal seleccionou-se para o estudo as três "Primeiras Cartas I" (correspondentes, na hipótese formulada, à apresentação), e as três "Cartas Últimas" (correspondentes

à despedida); saliente-se que as três últimas cartas são na verdade cinco textos, uma vez que a autora que escreve a primeira das últimas cartas "ensaia" a despedida por duas vezes, mas só o faz à terceira. Note-se que se supôs que as três cartas que constituem a "Primeira Carta Última" são da mesma autora.

No entanto, não disputamos que outros possam pensar que este triplo texto seja a despedida das três autoras, e que duas delas, ainda insatisfeitas, tenham depois juntado uma espécie de post-scriptum (ou, se quisermos encará-lo como testamento, um codicilo).

Para além destes oito textos foram aleatoriamente escolhidos mais quatro; para tal recorreu-se à escolha aleatória das suas páginas. Os textos escolhidos foram:

- A Paz

- ***Lamento da Mariana Alcoforado para Dona Brites***
- ***Monólogo de uma mulher chamada Maria, com sua patroa***
- ***Carta de um escriturário, em África, para sua mulher de nome Mariana a viver em Lisboa.***

A estes textos acrescentou-se ainda:

- Redacção de uma rapariga de nome Maria Adélia nascida no Carvalhal e educada num asilo religioso em Beja,

por ser um dos textos preferidos da autora deste trabalho.

Foi para os treze textos atrás enunciados que se tentou fazer uma atribuição de autoria. Para isso estudou-se os trinta e cinco blocos de texto com autoria conhecida e tentou-se identificar e definir as variáveis que iriam, talvez, permitir estabelecer a autoria dos treze textos com autoria desconhecida.

Depois de escolhidos os textos de autoria conhecida e desconhecida a utilizar, o passo seguinte consistiu na escolha e quantificação das variáveis. O porquê da escolha das variáveis vai ser explicada no decorrer da apresentação do trabalho, à medida que estas forem sendo estudadas.

Quanto ao processo de quantificação, refira-se apenas que, embora pareça um processo simples e rápido (resume-se a contar o número de vezes que determinada palavra ocorre ao longo de um bloco), se relevou um processo extremamente difícil e moroso, pois ao longo de todo este trabalho nenhum software específico foi utilizado. Ou seja, todo o processo de contagem, nomeadamente das palavras, foi efectuado apenas com a ajuda do processador de texto Word.

3. Tentativa de estudo de algumas variáveis discriminativas das autoras

Nesta secção apresentam-se as variáveis estudadas (ou que se tentaram estudar) e as conclusões a que o seu estudo conduziu.

É dedicado um parágrafo a cada variável que se estudou, onde se apresenta a motivação que levou à sua escolha, se expõe, de maneira breve, o modo como foi efectuada a recolha dos dados, e, finalmente se apresenta o estudo efectuado e as conclusões a que o mesmo conduziu.

3.1. As palavras

Quando alguém se propõe fazer um estudo deste tipo e se depara com a necessidade de definir variáveis, pensa de imediato na frequência das palavras. Não só por ser natural, mas também porque, deste modo, dispõe de um número significativo de variáveis sobre o qual se pode debruçar; note-se que um texto pode ter centenas ou mesmo milhares de palavras, logo o número de possíveis variáveis a estudar é vasto.

Mas será que todas as palavras num texto têm interesse, ou podem-se “ignorar” algumas?

3.1.1. Separação das palavras em contextuais e não contextuais

Sempre que alguém, escritor ou não, escreve um texto, grande parte das palavras utilizadas são contextuais, i.e., palavras que dependem do assunto sobre o qual se escreve. Por exemplo, se um texto versar sobre a saúde, o vocabulário nele encontrado vai estar intimamente ligado à área em questão; é por isso natural que em tal texto as palavras saúde, médico e hospital, por exemplo, tenham uma frequência elevada. Já se o texto for sobre o amor, é natural que as palavras saúde, médico e

hospital tenham uma frequência muito baixa ou mesmo nula e as palavras amor, paixão e carinho tenham, por sua vez, uma frequência elevada.

Dos exemplos anteriores conclui-se que, para este estudo, nem todas as palavras interessam, ou seja, as palavras que dependam do contexto — palavras contextuais — não interessam, pois o facto de não serem utilizadas num determinado texto nada diz sobre a sua frequência de utilização pelo autor — tudo está dependente do contexto. Assim, as palavras que têm interesse para este tipo de estudos são palavras que independentemente do assunto, i.e., do contexto, são sempre utilizadas — palavras não contextuais. A esta classe de palavras pertencem vocábulos do tipo: “e”, “para”, “sim”, “não”, etc., e, de um modo geral, pronomes, conjunções e advérbios.

Os textos analisados em primeiro lugar, e por motivos óbvios, foram os de autoria conhecida. Assim, para estes textos, começou-se por separar as palavras que neles figuravam em dois grandes grupos: as contextuais e as não contextuais. Note-se que, no entanto, esta separação não é tão linear como aparenta, pois palavras existem que são difíceis de classificar. As palavras "branco", "criança" e "homem", por exemplo, são ou não palavras contextuais? Neste estudo foram classificadas como sendo não contextuais.

Em apêndice apresenta-se a separação, das palavras em contextuais e não contextuais, efectuada para todos os blocos de texto conhecido de cada autora. Devido ao facto da separação das palavras em contextuais e não contextuais ser um tanto ou quanto subjectiva, a probabilidade de existir quem não concorde com a classificação efectuada é grande, mas, quando foi efectuada pareceu-nos ser a mais indicada.

Apresenta-se, na tabela seguinte, o número médio de palavras (diferentes), dos dois tipos, que foram utilizadas em cada bloco de texto pelas autoras.

	Isabel	**Teresa**	**Velho**
Palavras não contextuais	122.92	121	134.92
Palavras contextuais	465.08	397.18	442.67

Tabela 3: Média de palavras contextuais e não contextuais utilizadas por autora

Analisando a tabela anterior, a ideia que esta transmite é que o número de palavras não contextuais utilizado é pequeno comparativamente ao número de palavras contextuais; a Maria Teresa e a Maria Velho utilizam três vezes mais palavras contextuais do que não contextuais, enquanto, a Maria Isabel utiliza cerca de quatro vezes mais. No entanto, são as palavras não contextuais que apresentam uma frequência mais elevada. De seguida, apresenta-se a frequência total das palavras não contextuais utilizadas por cada autora.

	Isabel	**Teresa**	**Velho**
Frequência total das palavras não contextuais	9127	8543	9939
Percentagem	50.77%	51.78%	55.22%

Tabela 4: Frequência total das palavras não contextuais por cada autora

Ou seja, um pequeno número de palavras representa cerca de 50% dos dados em estudo, indo, no caso da Maria Velho, até aos 55%. Tal facto é compreensível se se tiver em consideração que, mesmo quando se escreve um "texto corriqueiro", vocábulos como: "a", "e", e "que" são usados com bastante frequência, e é este tipo de palavras que forma o conjunto das não contextuais.

Depois de efectuada a separação em palavras contextuais e não contextuais, as primeiras foram "eliminadas" do estudo, pelo que, no que se segue, apenas se utilizaram as palavras não contextuais.

O problema que se colocou a seguir foi o de decidir se as palavras não contextuais, utilizadas nos textos de autoria conhecida, se deveriam utilizar todas, ou se entre estas existiam palavras com um maior poder discriminativo.

3.1.2. Selecção das palavras não contextuais a estudar

Mosteller e Wallace, [8], chamam a atenção para o facto de palavras com grande variabilidade não serem, de modo geral, boas discriminadoras e apresentam um índice, Z, como medida do poder discriminador de uma palavra. Uma vez que Mosteller e Wallace trabalham apenas com dois autores, foi uma generalização do seu índice Z que foi utilizada para medir o poder de discriminação de uma palavra.

Considere-se então o terno ordenado (x, y ,w), onde x, y e w indicam que determinada palavra ocorreu em x textos da Maria Isabel, y textos da Maria Teresa e w textos da Maria Velho (textos de autoria conhecida). Do conjunto das palavras não contextuais foram retidas as que apresentavam um "score" do tipo $|x-y|\geq 2$, ou $|x-w|\geq 2$ ou ainda $|y-w|\geq 2$, i.e., foram retidas as palavras que, para uma autora, ocorreram em pelo menos mais dois textos do que para as restantes. Para cada palavra seleccionada, calculou-se o índice

$$Z = \frac{(x-y)^2}{x+y} + \frac{(x-w)^2}{x+w} + \frac{(y-w)^2}{y+w}$$

como medida do poder discriminador dessa palavras. O índice atrás apresentado pode ser entendido como a "estatística de teste" do Qui-Quadrado.

Depois de calculado o índice anterior retiveram-se as palavras para as quais $Z\geq 8$, onde 8 é o resultado do arredondamento de 7.81, valor do quantil de probabilidade $(1-\alpha)=0.95$ do Qui-Quadrado com três graus de liberdade; ou seja, pode-se dizer que se efectua o "teste do Qui-Quadrado" e que apenas se retiveram as palavras para as quais as três autoras apresentavam frequência de utilização significativamente diferente.

Chegou-se, deste modo, à lista de trinta e uma palavras que a seguir se apresenta.

Palavra	(x, y, w)	Z
Através	(0, 0, 9)	18
Bom	(10, 5, 0)	16.67
Breve	(0, 5, 2)	8.29
Certo	(3, 9, 0)	15
Cheia(s)	(5, 0, 1)	8.67
Comum(s)	(5, 2, 0)	8.29
Defronte	(4, 0, 0)	8
Dessa(s)	(4, 0, 0)	8
Donde	(0, 4, 4)	8
Duma(s)	(11, 5, 2)	9.77
Enorme(s)	(1, 3, 10)	12.13
Gente(s)	(9, 8, 1)	11.9
Jamais	(0, 7, 1)	12.5
Maior(es)	(0, 7, 1)	12.5
Maneira(s)	(3, 0, 10)	16.77
Menino(s)	(6, 4, 0)	10.4
Muita(s)	(9, 6, 0)	15.6
Naquela(s)	(0, 4, 5)	9.11
Nele(s)	(0, 5, 2)	8.29
Ninguém	(10, 2, 2)	10.67
Nova(s)	(9, 8, 1)	11.90
Perto	(2, 6, 10)	8.33
Pessoa(s)	(10, 0, 9)	19.05
Pessoal(ais)	(6, 3, 0)	10
Porém	(0, 8, 5)	13.69
Quanto(s)	(0, 5, 2)	8.29
Senhora(s)	(6, 2, 0)	10
Somente	(0, 6, 7)	13.08
Tanta(s)	(12, 2, 2)	14.29
Todavia	(0, 0, 11)	22
Total	(0, 0, 4)	8

Tabela 5: Palavras não contextuais seleccionadas

De notar que o teste efectuado se revelou bastante eficaz, pois a maioria das palavras seleccionadas apresentam frequências de utilização muito diferentes para as três autoras, existindo entre as trinta e uma palavras seleccionadas quatro que se

destacam por a elas estarem associadas ternos do tipo: (x, 0, 0) ou (0, 0, w). Ou seja, o "teste" efectuado conseguiu identificar um tipo de palavras muito importante em estudos de autoria, designado por palavras marca, e que são palavras identificativas de um dos autores, i.e., palavras que apenas um dos autores utiliza.

Foi para todas as palavras que constam da lista anterior que o estudo prosseguiu.

3.1.3. Estudos efectuados para as palavras não contextuais seleccionadas

Para as palavras seleccionadas pelo processo anteriormente descrito calculou-se a frequência absoluta em cada um dos textos das três autoras. Depois, e seguindo também uma sugestão de Mosteller e Wallace, a partir das frequências encontradas calculou-se o número de palavras esperadas em blocos de mil palavras; ou seja, para uma dada palavra e utilizando a sua frequência, calculou-se o número de vezes que essa palavra tenderia a aparecer num bloco com mil palavras — i.e., a sua permilagem. Mosteller e Wallace justificam o seu conselho com o argumento de que depois será mais fácil comparar as frequências observadas nos textos de autoria conhecida com as observadas nos textos de autoria desconhecida. Dizer que, por exemplo, a frequência do vocábulo "não" é vinte não é muito esclarecedor, pois vinte poderá ser considerada uma frequência alta ou baixa conforme o texto em análise seja grande ou pequeno. Considerar a permilagem das frequências das palavras, quer nos textos conhecidos quer nos "desconhecidos", facilita a análise, pois a "unidade" é a mesma.

De seguida calculou-se a média, a mediana e o desvio-padrão para as amostras de permilagens. Os resultados obtidos são apresentados na tabela que se segue.

Palavra	Isabel			Teresa			Velho		
	Média	**Med**	**D. P.**	**Média**	**Med**	**D. P.**	**Média**	**Med**	**D. P.**
Através	-	-	-	1.91	2	1.22	-	-	-
Bom	0.56	0.67	0.62	-	-	-	0.28	0	0.45
Breve	-	-	-	0.12	0	0.27	0.45	0	0.77
Certo	0.17	0	0.30	-	-	-	1.56	1.67	1.28
Cheia(s)	0.55	0	0.84	0.06	0	0.20	-	-	-
Comum(s)	0.28	0	0.35	-	-	-	0.17	0	0.41
Defronte	0.22	0	0.33	-	-	-	-	-	-
Dessa(s)	0.28	0	0.45	-	-	-	-	-	-
Donde	-	-	-	-	-	-	0.22	0	0.33
Duma(s)	0.89	0.67	0.43	0.12	0	0.27	0.45	0	0.66
Enorme(s)	0.06	0	0.19	1.58	1.33	1.04	0.17	0	0.41
Gente(s)	1.78	0.67	2.13	0.06	0	0.20	0.83	0.67	0.81
Jamais	-	-	-	0.06	0	0.20	0.89	0.67	1.18
Maior(es)	-	-	-	0.06	0	0.20	0.56	0.67	0.56
Maneira(s)	0.17	0	0.30	0.85	0.67	0.43	-	-	-
Menino(s)	0.95	0.34	1.29	-	-	-	0.56	0	0.89
Muita(s)	0.61	0.67	0.44	-	-	-	0.45	0.34	0.52
Naquela(s)	-	-	-	0.36	0	0.46	0.45	0	0.77
Nele(s)	-	-	-	0.12	0	0.27	0.39	0	0.53
Ninguém	1.17	1	0.99	0.12	0	0.46	0.11	0	0.26
Nova(s)	1.11	1.33	0.77	0.06	0	0.20	0.67	0.67	0.57
Perto	0.11	0	0.26	1.76	1.33	1.50	0.50	0.34	0.76
Pessoa(s)	2.67	2.67	1.95	1.09	1.33	0.80	-	-	-
Pessoal(ais)	0.39	0.34	0.45	-	-	-	0.28	0	0.60
Porém	-	-	-	0.55	0	0.72	1.50	2	1.28
Quanto(s)	-	-	-	0.36	0	0.81	0.5	0	0.70
Senhora(s)	0.78	0.34	1.09	-	-	-	0.28	0	0.78
Somente	-	-	-	0.43	0.67	0.34	0.34	0.34	0.35
Tanta(s)	1.28	1	0.78	0.12	0	0.27	0.11	0	0.26
Todavia	-	-	-	0.67	0.67	0.99	-	-	-
Total	-	-	-	0.42	0	0.68	-	-	-

Tabela 6 : Média, mediana e desvio-padrão para a frequência das palavras seleccionadas

Da análise da tabela anterior ressaltam dois factos. Um tem a ver com diferença entre média e a mediana que, de um modo geral, não é significativa; como tal resolveu trabalhar-se apenas com a média. O outro diz respeito ao valor do desvio-padrão associado a cada palavra, quase todas as palavras apresentam um desvio-

padrão inferior a dois, ou seja, as palavras seleccionadas são, em princípio, boas discriminadores das autoras. Por outro lado, o facto das palavras seleccionadas apresentarem um desvio pequeno vem, de algum modo, validar o índice Z utilizado como medida do poder discriminador das palavras.

Para se poder prosseguir com o estudo tornou-se necessário conhecer a frequência das palavras seleccionadas nos textos "desconhecidos". Assim, para os textos de autoria desconhecida foi verificar-se se alguma das palavras seleccionadas neles ocorria e, caso afirmativo a sua frequência.

O primeiro "contratempo" surgiu neste ponto. Constatou-se que existiam textos onde nenhuma das palavras seleccionadas ocorria, ou ocorria um número reduzido destas e, neste caso, a frequência de ocorrência de cada uma era, por vezes, bastante baixa.

Mesmo assim havia que analisar os dados disponíveis. Deste modo, havia que, em primeiro lugar, converter a frequência das palavras, que ocorreram nos textos "desconhecidos", na frequência esperada em blocos de mil palavras. De seguida, havia que escolher a ferramenta estatística a empregar para se comparar os dados e tirar, se possível, algumas conclusões. Mas, novo "contratempo" ocorreu.

Apesar do número quase infindável de técnicas estatísticas existentes, nenhuma delas pareceu adequada para efectuar o estudo anterior. Que técnica estatística se deve utilizar quando o que se pretende é verificar de que população um dado valor provém? Testes de hipóteses estavam fora de questão pois vários problemas se colocavam ao seu uso, tais como a normalidade dos dados. Testes de ajustamento também não pareceram adequados; que teste se deve utilizar para verificar se um único valor provém de uma população? Perante este quadro resolveu-se recorrer à técnica estatística mais velha e talvez mais utilizada, e comparar o valor encontrado para o texto "desconhecido" com a média calculada para os textos conhecidos, e dizer que o valor provém da população que tem valor médio mais "próximo" desse valor. Ou seja, perante um valor de um texto "desconhecido" se esse valor se "aproximar" mais da média calculada para os textos da Maria Isabel, por exemplo, então dir-se-á que provavelmente o texto desconhecido é da Maria Isabel.

É claro que o uso deste procedimento é polémico e muitas e variadas questões se colocam: desde logo o facto de se basear apenas na média. Talvez as suas únicas vantagens sejam a simplicidade e a aplicabilidade, já que pode ser sempre utilizado e a simplicidade é evidente. As escolhas, como já foi referido, não eram muitas, pelo que se deixa ao cuidado do leitor os julgamentos que achar oportunos.

Apresentam-se, de seguida, as conclusões retiradas à custa do procedimento anteriormente descrito. Para melhor visualização dos resultados apresenta-se uma tabela para cada um dos textos estudados. Para a "Primeira Carta I" não é apresentada nenhuma tabela por não existir neste texto nenhuma das palavras seleccionadas; note-se no entanto, que dos textos seleccionados, este é um dos mais pequenos com apenas cento e setenta e seis palavras.

Palavras	**Texto desconhecido**	**Isabel**	**Teresa**	**Velho**	**Tendência**
Através	3.95	-	1.91	-	Teresa
Certo	1.98	0.17	-	1.56	***Velho***
Jamais	3.95	-	0.06	0.89	Velho
Maior(es)	3.95	-	0.06	0.56	Velho
Maneira(s)	1.98	0.17	0.85	-	Teresa
Total	1.98	-	0.68	-	Teresa
					Teresa/Velho

Tabela 7 : Primeira Carta II

Palavras	**Texto desconhecido**	**Isabel**	**Teresa**	**Velho**	**Tendência**
Porém	6.69	-	0.55	1.5	Velho
					Velho

Tabela 8: Primeira Carta III

Palavras	Texto desconhecido	Isabel	Teresa	Velho	Tendência
Bom	1.65	0.56	-	0.28	Isabel
Gente(s)	1.65	1.78	0.06	0.83	Isabel
Maior(es)	1.65	-	0.06	0.56	Velho
Nele(s)	1.65	-	0.12	0.39	Velho
Ninguém	3.31	1.17	0.12	0.11	Isabel
					Isabel

Tabela 9: Primeira Carta Última - Parte 1

Palavras	Texto desconhecido	Isabel	Teresa	Velho	Tendência
Gente(s)	1.63	1.78	0.06	0.83	Isabel
Muita(s)	1.63	0.61	-	0.45	Isabel
Ninguém	3.27	1.17	0.412	0.11	Isabel
Quanto(s)	3.27	-	0.36	0.5	Velho
Somente	1.63	-	0.43	0.34	**Teresa**
					Isabel

Tabela 10: Primeira Carta Última - Parte 2

Palavras	Texto desconhecido	Isabel	Teresa	Velho	Tendência
Bom	1.05	0.56	-	0.28	Isabel
Certo	1.05	0.17	-	1.56	Velho
Dessa(s)	1.05	0.28	-	-	Isabel
Maneira(s)	2.09	0.17	0.85	-	Teresa
Muita(s)	1.05	0.61	-	0.45	Isabel
Ninguém	1.05	1.17	-	0.11	Isabel
Perto	1.05	0.11	1.76	0.50	Velho
					Isabel

Tabela 11: Primeira Carta Última – Parte 3

Palavras	Texto desconhecido	Isabel	Teresa	Velho	Tendência
Bom	2.80	0.56	-	0.28	Isabel
Certo	0.93	0.17	-	1.56	Velho
Cheia(s)	0.47	0.55	0.06	-	Isabel
Gente(s)	1.40	1.78	0.06	0.83	Isabel
Menino(s)	0.93	0.95	-	0.56	Isabel
Muita(s)	2.33	0.61	-	0.34	Isabel
Ninguém	0.47	1.17	0.12	0.11	Teresa/Velho
Nova(s)	0.47	1.18	0.06	0.67	Velho
Pessoa(s)	0.47	2.67	1.09	-	Teresa
Pessoal(ais)	0.47	0.39	-	0.28	Isabel
Quanto(s)	0.93	-	0.36	0.50	Velho
					Isabel

Tabela 12: Segunda Carta Última

Palavras	Texto desconhecido	Isabel	Teresa	Velho	Tendência
Jamais	5.62	-	0.06	0.89	Velho
Pessoa(s)	5.62	2.67	1.09	-	Isabel
Somente	5.62	-	0.43	0.34	Teresa
Tanta(s)	5.62	1.28	0.12	0.11	Isabel
					Isabel

Tabela 13: Terceira Carta Última

Palavras	Texto desconhecido	Isabel	Teresa	Velho	Tendência
Breve	1.89	-	0.12	0.45	Velho
Enorme(s)	3.79	0.06	1.58	0.17	Teresa
Nele(s)	1.89	-	0.12	0.39	Velho
Todavia	1.89	-	0.67	-	Teresa
					Teresa/Velho

Tabela 14: A Paz

Palavras	Texto desconhecido	Isabel	Teresa	Velho	Tendência
Jamais	2.24	-	0.06	0.89	Velho
Porém	3.36	-	0.72	1.5	Velho
Tanta(s)	1.12	1.28	0.12	0.11	Isabel
Todavia	1.12	-	0.67	-	Teresa
					Velho

Tabela 15: Lamento

Palavras	Texto desconhecido	Isabel	Teresa	Velho	Tendência
Bom	2.92	0.56	-	0.28	Isabel
Gente(s)	5.85	1.78	0.06	0.83	Isabel
Menino(s)	2.92	1.29	-	0.56	Isabel
Ninguém	1.46	1.17	0.12	0.11	Isabel
Pessoa(s)	1.46	2.67	1.09	-	Teresa
Senhora(s)	17.54	0.78	-	0.28	Isabel
Tanta(s)	1.46	1.28	0.12	0.11	Isabel
					Isabel

Tabela 16: Monólogo

Palavras	Texto desconhecido	Isabel	Teresa	Velho	Tendência
Breve	2.11	-	0.12	0.45	Velho
					Velho

Tabela 17: Escriturário

Palavras	Texto desconhecido	Isabel	Teresa	Velho	Tendência
Cheia(s)	0.74	0.55	0.06	-	Isabel
Dessa(s)	0.74	0.28	-	-	Isabel
Gente(s)	5.15	1.78	0.06	0.83	Isabel

Maneira(s)	0.74	0.17	0.85	-	Teresa
Muita(s)	1.47	0.61	-	0.45	Isabel
Ninguém	0.74	1.17	0.12	0.11	Isabel
Pessoa(s)	2.21	2.67	1.09	-	Isabel
					Isabel

Tabela 18: Tarefas

Da análise das tabelas anteriores, a primeira conclusão a retirar é que as autoras Maria Isabel e Maria Velho aparecem como as mais prováveis da maioria dos textos. Apesar de na maioria destes existirem valores que "apontam" na direcção da Maria Teresa, existem apenas dois ("Primeira Carta II" e a "Paz") onde autoria poderá ser discutida entre esta autora e a Maria Velho, já que, para os dois textos referidos, as conclusões apontam na direcção destas duas autoras.

Para sintetizar de uma forma mais clara os resultados obtidos, apresenta-se de seguida uma tabela com os textos estudados e as conclusões a que se chegou para cada um.

Texto	**Possível autor**
1ª Carta II	Teresa ou Velho
1ª Carta III	Velho
1ª Carta Última - parte 1	Isabel
1ª Carta Última - parte 2	Isabel
1ª Carta Última - parte 3	Isabel
2ª Carta Última	Isabel
3ª Carta Última	Isabel
Paz	Teresa ou Velho
Lamento	Velho
Monólogo	Isabel
Escriturário	Velho
Tarefas	Isabel

Tabela 19 : Primeira "atribuição" de autoria

Os resultados que constam da tabela anterior não são totalmente desencorajantes. Como foi dito no princípio deste trabalho, a Maria Teresa Horta é poetisa e partindo

do princípio que os poemas que constam nas Novas Cartas são seus, “é natural” que os textos em prosa sejam em maior número da autoria de Maria Isabel e de Maria Velho, facto que, a ser verdade, justificaria em parte os resultados anteriores.

Perante os resultados anteriores também não é de descartar a hipótese de uma das três autoras ter tido o trabalho de revisão dos textos, e deste modo ter “contaminado” os textos das colegas com o seu próprio estilo. Se tal aconteceu, as autoras que se afiguram como as revisoras mais prováveis são a Maria Isabel Barreno e a Maria Velho da Costa.

Chegados a este ponto e como as dúvidas eram cada vez mais, houve a necessidade de aferir, de alguma forma, a técnica utilizada. Pensou-se então em utilizar textos de autoria conhecida, diferentes dos até aqui usados, e aplicar a estes a mesma técnica para ver se esta conseguia “acertar” com a sua autoria. Resolveu-se então, para cada autora, considerar seis novos textos de autoria conhecida, e aplicar a estes o processo anteriormente descrito. Para a Maria Teresa Horta não foi possível arranjar novos textos, pois o conteúdo de “Ambas as Mãos Sobre o Corpo” foi utilizado na totalidade para construir os onze blocos de texto utilizados anteriormente, e, como já foi dito, não existe mais nenhum livro de prosa publicado por esta autora. Resolveu-se prosseguir com o teste apenas para a Maria Isabel e para a Maria Velho. Salienta-se que estes novos textos foram construídos de modo análogo aos restantes textos conhecidos, e têm a mesma dimensão, i. e., mil e quinhentas palavras.

Os resultados a que se chegou são apresentados nas tabelas seguintes.

Textos da Maria Isabel Barreno

Palavras	Bloco 1	Isabel	Velho	Tendência
Comum(s)	0.67	0.28	-	Isabel
Gente	0.67	1.78	0.83	Velho
				Isabel ou Velho

Tabela 20: Autoria atribuída ao novo bloco um

Palavras	Bloco 2	Isabel	Velho	Tendência
Pessoa(s)	1.33	2.67	-	Isabel
Pessoal(ais)	1.33	0.39	0.28	Isabel
Quanto(s)	0.67	-	0.5	Velho
Senhora(s)	0.67	0.78	0.28	Isabel
				Isabel

Tabela 21:Autoria atribuída ao novo bloco dois

Palavras	Bloco 3	Isabel	Velho	Tendência
Bom	0.67	0.56	0.28	Isabel
Breve	0.67	-	0.45	Velho
Dessa(s)	0.67	0.28	-	Isabel
Donde	0.67	-	0.22	Velho
Duma(s)	1.33	0.89	0.45	Isabel
Ninguém	0.67	1.17	0.11	Isabel
Nova(s)	1.33	1.18	0.67	Isabel
Pessoa(s)	1.33	2.67	-	Isabel
				Isabel

Tabela 22: Autoria atribuída ao novo bloco três

Palavras	Bloco 4	Isabel	Velho	Tendência
Bom	2.67	0.56	0.28	Isabel
Certo	0.67	0.17	1.56	Isabel
Duma(s)	1.33	0.89	0.45	Isabel
Gente(s)	0.67	1.78	0.11	Velho
Ninguém	0.67	1.17	0.11	Isabel
Nova(s)	0.67	1.18	0.67	Velho
				Isabel

Tabela 23:Autoria atribuída ao novo bloco quatro

Palavras	Bloco 5	Isabel	Velho	Tendência
Através	0.67	-	-	-
Duma(s)	0.67	0.89	0.45	Velho/Isabel
Gente(s)	2.67	1.78	0.83	Isabel
				Isabel

Tabela 24:Autoria atribuída ao novo bloco cinco

Palavras	Bloco 6	Isabel	Velho	Tendência
Cheia(s)	1.33	0.55	-	Isabel
Defronte	0.67	0.22	-	Isabel
Duma(s)	1.33	0.89	0.45	Isabel
Maior(es)	0.67	-	0.56	Velho
Menino(s)	0.67	0.95	0.56	Velho
Nova(s)	0.67	1.18	0.67	Velho
Pessoa(s)	3.33	2.67	-	Isabel
				Isabel

Tabela 25:Autoria atribuída ao novo bloco seis

Textos da Maria Velho da Costa

Palavras	**Bloco 1**	**Isabel**	**Velho**	**Tendência**
Bom	0.67	0.56	0.57	**Velho**

Tabela 26:Autoria atribuída ao novo bloco um

Palavras	**Bloco 4**	**Isabel**	**Velho**	**Tendência**
Através	1.33	-	-	-
Bom	1.33	0.56	0.56	Isabel/Velho
Quanto(s)	1.33	-	0.5	Velho
				Velho

Tabela 27:Autoria atribuída ao novo bloco quatro

Palavras	**Bloco 5**	**Isabel**	**Velho**	**Tendência**
Total	0.67	-	-	
				Inconclusivo

Tabela 28:Autoria atribuída ao novo bloco cinco

Analisando os resultados obtidos para o teste conclui-se, que este não funciona tão mal como se temia. Dos seis blocos de textos analisados para a Maria Isabel, a técnica utilizada apenas não permitiu chegar a nenhuma conclusão para o bloco número um; note-se, no entanto, que este bloco apenas contém duas das palavras seleccionadas o que não permitiu grandes estudos. Para os seis blocos de texto analisados para a Maria Velho, a primeira constatação a fazer reside no facto de nestes blocos existirem poucas das palavras seleccionadas; nos blocos dois, três e seis não ocorreu nenhuma das palavras seleccionadas e em dois dos restantes blocos apenas apareceu uma. No entanto, a frequência das palavras que ocorreram permitiu sempre, utilizando a técnica anterior, concluir que os textos eram da Maria Velho, excepto para o bloco cinco onde só ocorreu a palavra "total" que é uma palavra

marca da Maria Teresa. Ou seja, apesar das muitas reticências colocadas ao estudo efectuado, depois de verificados os resultados obtidos com o teste anterior, aumentou a confiança neste e nos resultados por ele alcançados.

3.2. Algumas palavras especiais

Antes de se prosseguir para o estudo de outras variáveis que não têm, directamente, que ver com a frequência das palavras utilizadas, vão apresentar-se os estudos efectuados para seis palavras em particular. A justificação para a escolha destes vocábulos será apresentada com o seu estudo.

3.2.1. Utilização do vocábulo “Certo” como qualificativo ou determinativo

A palavra “certo” pode ser empregada numa frase com dois significados: como qualificativo ou como determinativo. Por exemplo, na frase “Tu estás certo.”, certo é usado como qualificativo, enquanto na frase “Certo dia ...” é usado como determinativo.

Foi assim investigar-se qual das formas era mais utilizada pelas autoras, ou seja, se as autoras utilizavam indistintamente as duas formas do vocábulo ou usavam mais uma do que outra. Para tal consideraram-se os textos de autoria conhecida, onde o vocábulo ocorria e foi estudar-se o modo como tal acontecia. Os resultados obtidos foram os seguintes:

Isabel	**certo**	Qualificativo	3
		Determinativo	0
Teresa	**certo**	Qualificativo	0
		Determinativo	0
Velho	**certo**	Qualificativo	10
		Determinativo	19

Tabela 29 : Utilização do vocábulo "certo" nos textos de autoria conhecida

Da análise da tabela anterior a primeira conclusão a retirar, é o facto de nos textos da Maria Teresa o vocábulo “certo” não ocorrer, ou seja, a ocorrência desta palavra não ajuda a identificar a autora. Já a Maria Isabel utilizou a palavra três vezes e todas como qualificativo; a Maria Velho é, das três autoras, a que usa a palavra com mais frequência (nos textos desta autora a palavra ocorreu vinte e nove vezes), e utiliza o vocábulo nas suas duas formas, embora o utilize mais vezes como determinativo — cerca de duas vezes mais nesta forma.

Para as amostras recolhidas para a Maria Isabel e para a Maria Velho calculou-se a média, a mediana e o desvio-padrão, para a proporção de ocorrências da palavra em mil.

	Isabel		**Velho**	
	Qualificativo	**Determinativo**	**Qualificativo**	**Determinativo**
Média	0.11	-	0.56	1.06
Mediana	0	-	0.67	0.34
D. Padrão	0.26	-	0.36	1.29

Tabela 30: Média, mediana e desvio-padrão para a frequência da palavra "certo”

Da tabela anterior constata-se que a Maria Isabel raramente utiliza a palavra “certo”, já que a mediana de frequência do vocábulo em mil palavras é exactamente zero. Espera-se, deste modo, que nos textos da Maria Isabel a palavra ocorra um número reduzido de vezes e quando ocorrer seja utilizada como qualificativo.

No caso de textos da Maria Velho espera-se, que num texto com mil palavras, o vocábulo seja utilizado com maior probabilidade como determinativo do que como qualificativo, pois a probabilidade da autora utilizar a forma determinativa do vocábulo é cerca de duas vezes maior do que a de utilizar a forma qualificativa. De notar ainda, que, para a forma determinativa, a mediana e a média diferem consideravelmente; tal facto permitiu concluir que, quando esta autora utiliza num texto a palavra “certo” como determinativo, existe alguma probabilidade de que no mesmo texto, volte a utilizar outra vez o vocábulo na mesma forma.

Para se poder prosseguir com o estudo houve que voltar aos textos de autoria desconhecida e verificar em quais deles a palavra “certo” ocorria, qual a forma em que era utilizada e com que frequência. Verificou-se que a palavra apenas ocorria em três dos textos em estudo: uma única vez nos textos “Primeira Carta II” e “Primeira Carta Última-Parte3”, em ambas como determinativo, e duas vezes na “Segunda Carta Última”, sendo, neste caso, utilizado uma das vezes como determinativo e outra como qualificativo.

A investigação deparou-se, neste caso, com “falta de dados”, uma vez que, apenas quatro ocorrências da palavra não permite grandes estudos, e como tal, não pareceu adequado tirar conclusões a partir do modo como a palavra “certo” foi utilizada. O que apenas parece razoável concluir, atendendo ao facto de nos textos da Maria Teresa não ter ocorrido o vocábulo, é que a autoria dos três textos anteriores talvez se deva discutir entre a Maria Isabel e a Maria Velho. De salientar que o texto “Segunda Carta Última”, é no aspecto da utilização da palavra “certo”, um típico texto da Maria Velho, uma vez que o vocábulo é utilizado nas suas duas formas, facto que não foi observado para as outras duas autoras.

3.2.2. Estudo da posição dos vocábulos “Pois” e ”Depois” nas frases

As palavras “pois” e “depois” podem ter duas posições distintas numa frase, já que podem ser utilizadas no início ou no meio desta. Tal facto pareceu ser mais uma forma de se poder distinguir as autoras, se houvesse uma que utilizasse os vocábulos apenas no início das frases ou apenas no meio das frases. Foi então investigar-se o local onde estes vocábulos ocorriam nas frases dos textos conhecidos e desconhecidos.

Começa-se por apresentar o estudo efectuado para a palavra “**pois**”.

A tabela seguinte apresenta os dados para os textos de autoria conhecida

Isabel	**Pois**	Início	0
		Meio	3
Teresa	**Pois**	Início	0
		Meio	1
Velho	**Pois**	Início	8
		Meio	36

Tabela 31: Posição do vocábulo "pois" nas frases dos textos de autoria conhecida

Analisando a tabela anterior conclui-se que a palavra "pois" é utilizada com maior frequência pela Maria Velho e no meio de frases; nos textos da Maria Teresa a palavra ocorreu apenas uma vez no meio de uma frase, enquanto, a Maria Isabel utilizou a palavra três vezes também no meio de frases. Deste modo, textos onde a palavra seja empregue com alguma frequência talvez pertençam à Maria Velho; de notar ainda, que esta autora foi a única que nos textos analisados utilizou "pois" no início de frases, embora, utilize o vocábulo com mais frequência no meio.

Como os dados recolhidos para Maria Isabel e para a Maria Teresa são "pouco significativos" apresenta-se, de seguida, algumas medidas calculadas para as amostras recolhidas para a Maria Velho, depois, de calculada a proporção do vocábulo em mil palavras.

	Início	**Meio**
Média	0.44	2.00
Mediana	0	2
D. Padrão	0.71	1.30

Tabela 32:Algumas medidas calculadas para as amostras da Maria Velho

Analisando a tabela anterior verifica-se que num texto de mil palavras, se a autora utilizar o vocábulo, utiliza-o com maior probabilidade no meio de frases; repare-se que, em termos médios, num texto de mil palavras a autora utiliza a palavra duas vezes no meio de frases e nunca no início.

Quanto à ocorrência de "pois" nos textos de autoria desconhecida, os resultados apresentam-se na tabela abaixo.

1ª carta I	**Início**	1
	Meio	0
1ª carta II	**Início**	0
	Meio	1
1ª carta III	**Início**	0
	Meio	1
1ª Última-parte1	**Início**	0
	Meio	2
1ª Última-parte2	**Início**	0
	Meio	2
1ª Última-parte3	**Início**	0
	Meio	5
2ª Última	**Início**	1
	Meio	2
3ª Última	**Início**	0
	Meio	2
Paz	**Início**	0
	Meio	2
Lamento	**Início**	1
	Meio	3
Monólogo	**Início**	1
	Meio	2
Escriturário	**Início**	1
	Meio	1
Tarefas	**Início**	0
	Meio	4

Tabela 33: Ocorrências da palavra "pois" nos textos de autoria desconhecida

Analisando a tabela anterior, e tomando apenas em consideração a frequência absoluta de ocorrência da palavra nos textos, três textos se destacam por terem

frequências “elevadas”. São eles: “Primeira Carta Última - Parte3”, “Lamento” e “Tarefas” com frequências superiores a três; tais frequências levaram a supor que estes textos são da autoria da Maria Velho. O facto de a palavra ocorrer muito poucas vezes nos textos de autoria conhecida analisados para as outras duas autoras, conjugado com a baixa frequência da palavra nos textos de autoria desconhecida não permite ir mais longe na conclusão. A técnica até aqui utilizada, de comparar a média com o valor “observado” para o texto “desconhecido”, levou à conclusão de que todos os textos pertenciam à Maria Velho, o que parece ser pouco provável. Está-se perante um caso que nem a técnica elementar até aqui utilizada é útil, não se vislumbrando alternativas.

Apresentam-se, de seguida, os resultados obtidos para a palavra **“depois”.**

Isabel	Início	1
	Meio	27
Teresa	Início	5
	Meio	16
Velho	Início	7
	Meio	28

Tabela 34: Posição do vocábulo “depois” nas frases dos textos de autoria conhecida

Analisando a tabela anterior verifica-se que a Maria Velho é a autora que mais utiliza o vocábulo em questão, não se verificando, no entanto, uma discrepância tão grande na utilização do vocábulo pelas três autoras, como a que ocorria para a palavra “pois”. As três autoras utilizam a maior parte das vezes o vocábulo no meio das frases; porém, a Maria Teresa e a Maria Velho utilizam, com alguma frequência, o vocábulo no início de frases; em doze blocos de texto analisados apenas em um, e uma única vez, a Maria Isabel usou o vocábulo no início de uma frase. De salientar que, quando a Maria Teresa utiliza num texto “depois” no início de frase muitas vezes volta a utilizar, no mesmo texto, o vocábulo nessa posição, enquanto, nos textos da Maria Velho tal não se verifica.

Para melhor se compreender as amostras recolhidas calcularam-se as proporções em mil palavras e depois a média, a mediana e o desvio padrão. Os valores das estatísticas calculadas apresentam-se na tabela seguinte:

Autora	Posição	Média	Mediana	Desvio Padrão
Isabel	Início	0.06	0	0.19
	Meio	1.50	1	1.18
Teresa	Início	0.28	0	0.66
	Meio	0.89	0.67	0.87
Velho	Início	0.39	0.34	0.45
	Meio	1.56	1.67	1.34

Tabela 35: Algumas medidas calculadas para as três amostras

Em termos de valor médio, a Maria Velho é a que utiliza mais vezes a palavra "depois" no meio e no início de frases; saliente-se que as amostras da Maria Isabel e da Maria Velho, para o caso da utilização da palavra no meio de frases, apresentam a maior dispersão, enquanto, no caso da palavra ser utilizada no início da frase, são as amostras da Maria Teresa e da Maria Velho que apresentam maior dispersão.

Apresenta-se de seguida a frequência e a posição onde ocorreu o vocábulo "depois" nos textos de autoria desconhecida, assim como as respectivas permilagens.

Textos	Posição	Freq	Prop.
1ª Carta III	Início	0	-
	Meio	1	3.34
1ª Carta última - Parte 3	Início	0	-
	Meio	2	2.9
2ª Carta Última	Início	2	0.93
	Meio	3	1.40
3º Carta Última	Início	0	-
	Meio	1	5.62
Lamento	Início	0	-
	Meio	1	1.12
Monólogo	Início	0	-
	Meio	4	5.85
Tarefas	Início	1	0.74
	Meio	1	0.74

Tabela 36: Posição do vocábulo "depois" nas frases dos textos de autoria desconhecida

Comparando o valor das proporções anteriores com os valores médios calculados para os textos de autoria conhecida concluiu-se que:

Texto	Tendência
1ª Carta III	Velho
1ª Carta - Parte 3	Velho
2ª Carta Última	Isabel
3º Carta Última	Velho
Lamento	Teresa
Monólogo	Velho
Tarefas	Teresa

Tabela 37: Atribuição de autoria

Note-se que o texto "Terceira Carta Última" é muito pequeno (contém apenas cento e setenta e oito palavras) pelo que o valor da proporção encontrado poderá

estar enviesado. De notar ainda que a proporção 1.12 encontrada para o texto "Lamento" não é muito diferente do valor médio da Maria Isabel (1.50) pelo que o texto pode ser perfeitamente desta autora; aliás, o valor da mediana encontrado para esta autora vem reforçar o que atrás foi dito.

O estudo efectuado sobre a posição que as palavras "pois" e "depois" ocupam nas frases, conduziu a conclusões substancialmente diferentes, sob o ponto de vista da utilidade do estudo. Enquanto o estudo efectuado para o vocábulo "pois" não permitiu grandes conclusões, devido ao facto de ser um vocábulo pouco utilizado por duas das autoras e por ter uma frequência de ocorrência baixa nos textos de autoria desconhecida, já o estudo efectuado para o vocábulo "depois" permitiu a chegada a algumas conclusões. Tal ficou a dever-se, em parte, ao facto de o vocábulo ter uma frequência elevada tanto nos textos conhecidos como nos textos "desconhecidos".

3.2.3. Frequência de utilização dos vocábulos "de um/dum" e "de uma/duma"

O estudo da utilização dos vocábulos "de um" e "de uma" por oposição aos vocábulos "dum" e "duma", foi motivado pela constatação de que estes últimos são, de modo geral, menos utilizados na língua Portuguesa que os primeiros, e que uma das três autoras utilizava com alguma frequência os vocábulos "dum" e "duma", contrariamente ao que seria de esperar.

Assim, mais uma vez, recorreu-se aos textos conhecidos para estudar a frequência de utilização destes vocábulos pelas autoras. Os dados recolhidos nestes textos apresentam-se na tabela que se segue.

Autora	Vocábulo	Freq.
Isabel	De um	-
	De uma	-
	Dum	23
	Duma	14
Teresa	De um	26
	De uma	21
	Dum	1
	Duma	1
Velho	De um	14
	De uma	15
	Dum	9
	Duma	7

Tabela 38 : Frequência de utilização dos vocábulos “de um/dum” e “de uma/duma" nos textos conhecidos

Numa primeira análise da tabela anterior pode-se concluir, que nos doze textos da Isabel Barreno não se encontraram as expressões “de um” e “de uma” utilizando sempre, esta autora, as expressões “dum” e “duma”; já a Maria Teresa Horta utiliza as expressões “de um” e “de uma” com bastante frequência; nos onze textos desta autora apenas num deles apareceram as expressões “dum” e “duma”, e apenas uma vez cada. A Maria Velho é das três autoras a que utiliza com alguma frequência as quatro expressões, embora utilize as expressões “de um” e “de uma” com mais frequência.

Antes de se prosseguir com o estudo converteu-se as frequências observadas em proporções por mil palavras e de seguida calculou-se a média, a mediana e o desvio-padrão. Os resultados obtidos foram os seguintes:

Autora	Vocábulo	Média	Mediana	Desvio Padrão
Isabel	Dum	1.28	1	1.35
	Duma	0.78	0.67	0.48
Teresa	De um	1.45	0.67	1.75
	De uma	1.17	1.33	1.03
	Dum	0.06	0	0.19
	Duma	0.06	0	0.19
Velho	De um	0.78	0	1.10
	De uma	0.83	0.67	0.99
	Dum	0.50	0.67	0.50
	Duma	0.39	0	0.66

Tabela 39: Algumas medidas calculadas para a frequência dos vocábulos

Os resultados apresentados na tabela anterior vêm confirmar o que já tinha sido dito anteriormente. Note-se que as amostras recolhidas nos textos da Maria Isabel e nos textos da Maria Teresa relativamente ao uso da expressão “dum” e “de um”, respectivamente, são as que apresentam uma maior dispersão; ou seja, há textos onde as autoras usam as expressões com alguma frequência enquanto noutros essa frequência é mais baixa.

Nos textos de autoria desconhecida, encontraram-se as expressões em estudo em apenas cinco deles. A tabela abaixo apresenta os dados obtidos:

1ª Carta II	De uma	1
Monólogo	De uma	1
Escriturário	De um	1
	De uma	1
2ª carta última	De um	1
	De uma	1
	Dum	1

Tabela 40: Frequência de utilização dos vocábulos “de um/dum” e “de uma/duma” nos textos desconhecidos

Mais uma vez é muito difícil concluir sobre a autoria dos textos a partir destes dados. A única conclusão possível baseia-se no facto de a Maria Isabel nunca ter utilizado, nos textos de autoria conhecida, as expressões “de um” e “de uma” e como tal é pouco provável que tenha sido esta autora a escrever os textos anteriores; com as restantes autoras utilizam as quatro expressões, a autoria dos textos talvez deva ser discutida entre elas. De salientar, no entanto, que a Maria Teresa é das três autoras a que utiliza as expressões “de um” e “de uma” com mais frequência e como tal “aposta-se” mais nela como sendo a possível autora dos textos.

3.2.4. Frequência de utilização do vocábulo “Não”

Por ser uma palavra muito importante na língua Portuguesa e por ser uma das palavras com maior frequência de utilização nos textos, decidiu-se estudar a frequência de utilização da palavra “não”.

Em “O Português Fundamental”, o Professor Lindley Cintra estuda a frequência dos vocábulos mais utilizados na língua Portuguesa falada e escrita e neste estudo o vocábulo “não” aparece como sendo o terceiro vocábulo mais utilizado na língua Portuguesa.

A tabela abaixo apresenta a média, a mediana e o desvio-padrão calculados para as amostras de permilagens, obtidas a partir da frequência da expressão nos textos de autoria conhecida.

	Isabel	**Teresa**	**Velho**
Média	9.94	6.06	14.83
Mediana	10.33	6	13.67
Desvio Padrão	3.27	3.12	7.52

Tabela 41: Medidas calculadas para a frequência de utilização da palavra “não” nos textos conhecidos

Da análise da tabela anterior pode-se concluir, que a Maria Velho é a que utiliza mais vezes a palavra “não”, sendo a Maria Teresa a que utiliza menos vezes este vocábulo. Verifica-se ainda que a Maria Velho tanto utiliza o vocábulo com muita frequência como o utiliza com pouca frequência, sendo por isso, a autora para a qual a amostra recolhida apresenta uma maior dispersão. De notar ainda, que para as três autoras a diferença entre a média e a mediana não é muito significativa.

Para os textos de autoria desconhecida obtiveram-se as seguintes proporções por mil palavras:

Textos	Proporção
1ª Carta I	34.09
1ª Carta II	27.67
1ª Carta III	16.72
1ª Carta Última – Parte 1	38.01
1ª Carta Última – Parte 2	17.97
1ª Carta Última – Parte 3	20.92
2ª Carta Última	20.97
3ª Carta Última	11.24
Paz	5.68
Lamento	8.97
Monólogo	26.31
Escriturário	8.42
Tarefas	16.19

Tabela 42: Proporções calculadas a partir da frequência obtida nos textos desconhecidos

Se se compararem as proporções anteriores com os valores médios obtidos para os textos de autoria conhecida, conclui-se, que a palavra “não” não é uma boa discriminadora, e tal facto, talvez se verifique porque esta palavra apresenta uma variabilidade muito grande, especialmente para a Maria Velho.

Observando os valores da tabela verifica-se, que o valor da proporção encontrada para a “Primeira Carta III”, “Primeira Carta Última–Parte 2”, “Tarefas” e para a “Paz” são valores que indicam a Maria Velho como a possível autora dos três primeiros textos e a Maria Teresa como a possível autora do último. Os valores encontrados para os textos “Terceira Carta Última”, “Lamento” e “Escriturário” são

valores próximos daqueles que a Maria Isabel utiliza. Para os restantes textos os valores encontrados são demasiado grandes para se poder fazer qualquer julgamento.

Como já foi dito a frequência de uso da palavra “não” não é uma boa variável para discriminar as autoras e tal facto, talvez se deva, como já foi referido, à grande variabilidade que está associada ao uso deste vocábulo. Recorde-se que Mosteller e Wallace referem que palavras com grande variabilidade não devem ser utilizadas como discriminadoras, [8].

3.3. Comprimento de frase

Outras das variáveis estudadas neste trabalho, foi o comprimento de frase utilizado pelas autoras. Ao ler-se as obras em estudo, um dos aspectos que chamou a atenção, foi o facto de, pelo menos, duas das autoras, utilizarem frases extremamente longas. Por outro lado, o comprimento de frase aparece, em trabalhos desta natureza, como uma das variáveis que faz todo o sentido utilizar, uma vez que um dos aspectos que caracteriza o estilo de um autor, é, sem dúvida, a construção de frase que este utiliza, que pode ser caracterizada, entre outras coisas, pelo seu comprimento.

O comprimento de frase foi avaliado, como é óbvio, pelo número de palavras que constituem a frase; assim, uma frase longa é constituída por um número considerável de palavras, enquanto, que uma frase pequena é constituída por um número reduzido de palavras; pode acontecer que a frase seja constituída, apenas, por uma única palavra.

Antes de se prosseguir com a apresentação do estudo efectuado, convém explicitar o corpus utilizado. Como nos doze blocos de texto conhecido (onze, no caso da Maria Teresa), o número de frases era considerável, decidiu-se para o estudo, escolher cem frases de cada autora. A escolha de cem frases não se ficou a dever a qualquer motivo em especial; apenas, se considerou ser uma amostra de tamanho considerável de frases para análise, que permitia já uma análise cuidada. Outros teriam escolhido outro valor, mas, este pareceu-nos suficiente: nem muito pequeno, nem muito grande.

As cem frases, para cada autora, foram escolhidas aleatoriamente. Para cada um dos doze blocos (onze, no caso da Maria Teresa), escolheu-se aleatoriamente o número de frases a recolher em cada um, e, dentro de cada bloco seleccionou-se aleatoriamente quais as frases a utilizar. Por exemplo, se para o bloco seis da Maria Teresa, saísse quatro como o número de frases a escolher nesse bloco, o passo seguinte seria o de seleccionar, aleatoriamente, essas quatro frases; para tal geraram--se aleatoriamente quatro números, que corresponderiam aos números das frases a recolher; assim, se saísse o vinte, tal queria dizer, que a vigésima frase a contar do início iria ser utilizada. Note-se que um bloco de texto conhecido, não começa, ou acaba, necessariamente, com o começo, ou o final, de uma frase (tal deve-se ao modo como os blocos foram construídos), como é óbvio, nestes casos, a primeira e/ou a última frases foram eliminadas da escolha.

Em seguida, apresenta-se o estudo efectuado para o comprimento de frase. Para as amostras recolhidas calculou-se a média, a mediana, o mínimo, o máximo e o desvio padrão, encontrando-se os dados obtidos na tabela seguinte:

	Média	**Mediana**	**Mínimo**	**Máximo**	**Desvio Padrão**
Isabel	54	29,5	4	525	75.92
Teresa	32.27	22,5	1	184	33.01
Velho	31.74	24	2	175	30.02

Tabela 43: Algumas medidas para o comprimento de frase

Uma primeira conclusão, a retirar da análise da tabela anterior, tem a ver com o valor da média e da mediana, que são substancialmente diferentes. Tal facto resulta das amostras recolhidas serem assimétricas à direita (como à frente se mostra) e da média ser fortemente influenciada por valores extremos. Assim, e nesta subsecção, vai utilizar-se o valor da mediana, por ser uma das medidas de localização mais resistentes a valores extremos, para estabelecer comparações entre as autoras.

Quanto ao tamanho de frase utilizado pelas autoras conclui-se que a Maria Isabel é a que utiliza frases mais longas, se se utilizar a média como medida de comparação entre as autoras. A diferença entre o comprimento de frases utilizado pela Maria Isabel e pelas outras autoras é considerável; no entanto, em termos de mediana, essa diferença é quase esbatida. A amostra recolhida para a Maria Isabel é a que apresenta maior dispersão, concluindo-se, assim, que esta autora tanto pode utilizar frases extremamente longas como frases muito pequenas.

De seguida, apresentam-se as caixas-com-bigodes paralelas construídas para estes dados.

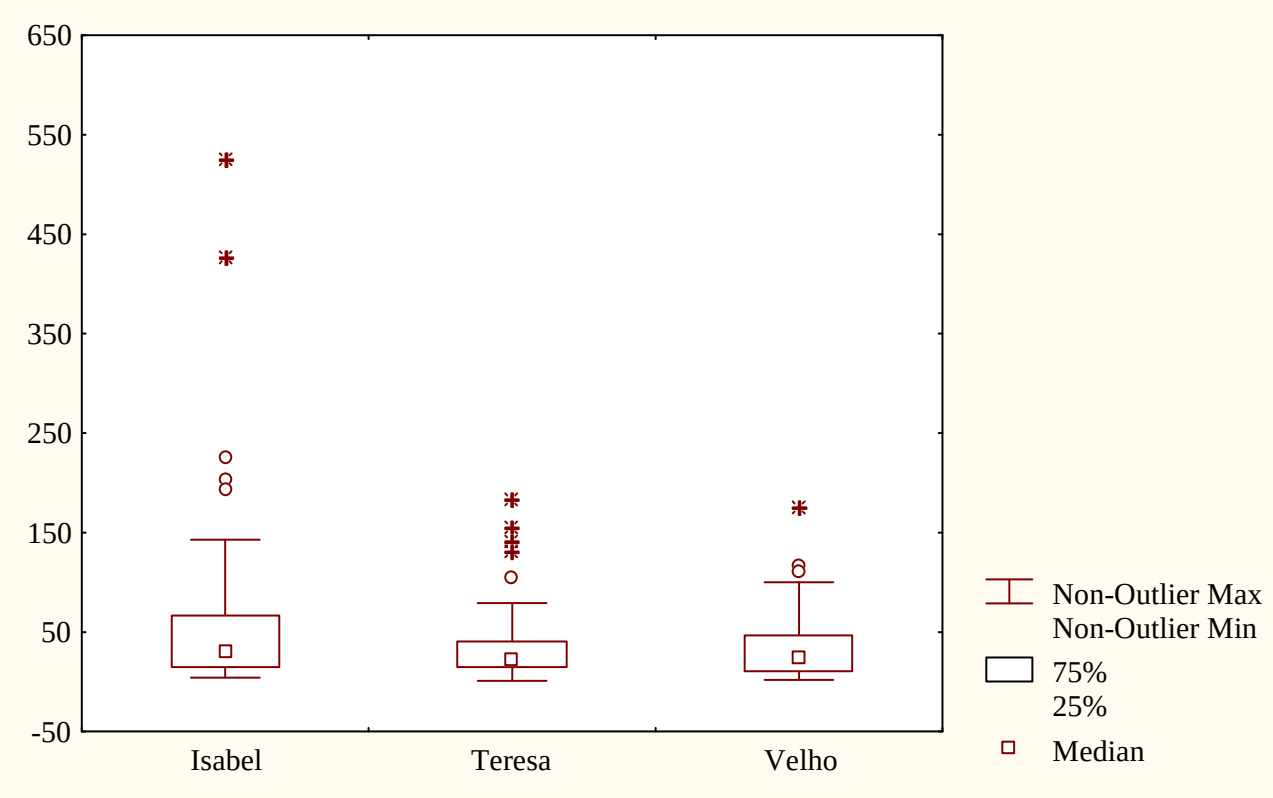

Figura 1: Caixas-com-bigodes paralelas para o comprimento de frase

Analisando a figura anterior, o primeiro aspecto que convém referir é a existência de um número considerável de outliers para as três colecções. As colecções apresentam-se pouco dispersas, com caudas direitas compridas e apresentando uma assimetria positiva.

Para melhor se poder analisar os dados, aplicou-se-lhes uma transformação potência, para tentar estabilizar a relação entre a dispersão e o nível. A recta ajustada aos pontos do gráfico dispersão-versus-nível apresentou um declive de b=2.413, pelo que, p=1-b=-1.413 é o expoente a utilizar na transformação potência, para estabilizar a dispersão. Como –1.413 é um valor próximo de –1.5, a transformação a efectuar em x é o recíproco da raiz quadrada de x ao cubo.

Ao desenhar este gráfico verificou-se, que a transformação anterior não só não melhorava a situação como a piorava. Experimentou-se ainda –2, como expoente da transformação mas, tal como para o caso p=-1.5, tal transformação revelou-se um perfeito desastre. Optou-se então por ignorar estas transformações e utilizar a transformação logaritmo. Apesar de não ser recomendada pelo gráfico dispersão-versus-nível, esta transformação revelou-se excelente. Note-se que o gráfico de dispersão-versus-nível é constituído, para o caso em estudo, por apenas três pontos, logo, basta uma pequena oscilação num deles para que a recta ajustada aos pontos sofra grandes perturbações. Este aspecto explica, em parte, o facto de a transformação sugerida pelo gráfico não ser a melhor.

Apresentam-se de seguida as caixas-com-bigodes paralelas para os dados logaritmizados.

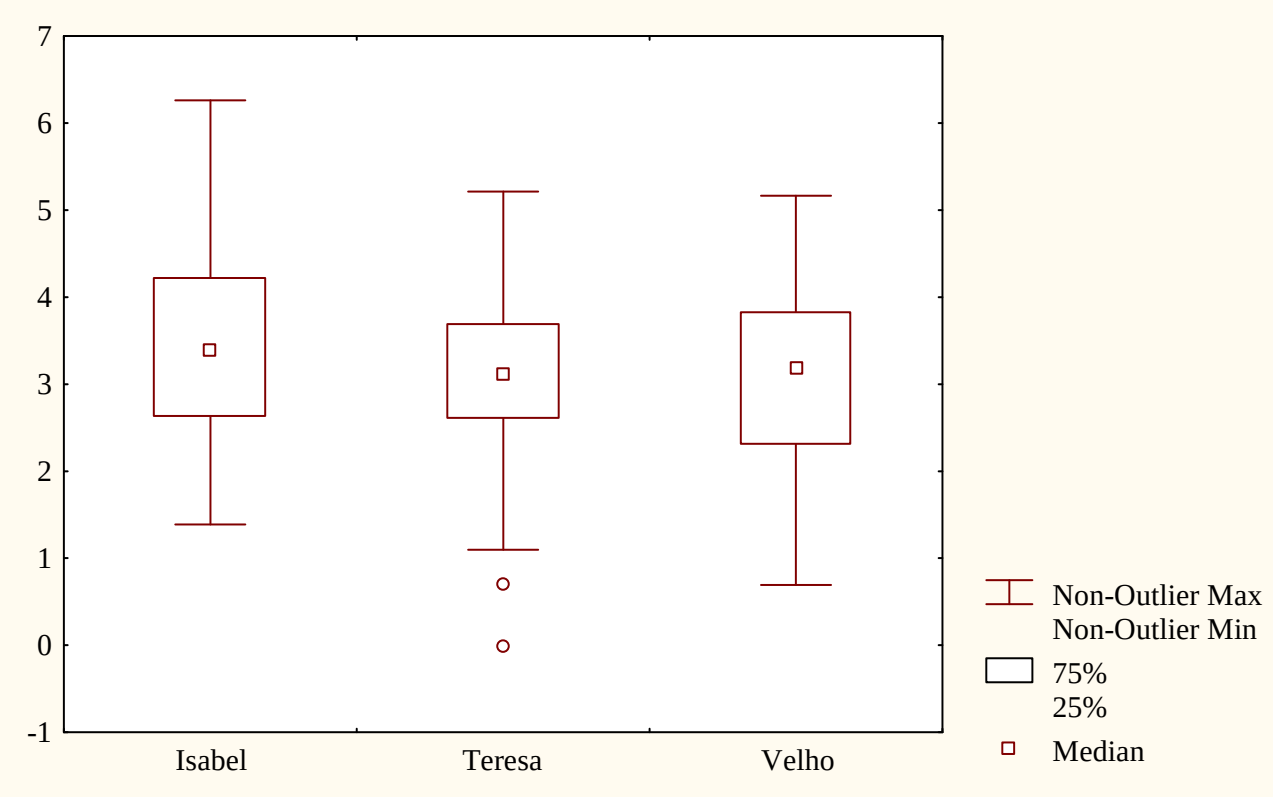

Figura 2: Caixas-com-bigodes paralelas para o logaritmo dos dados

A figura anterior permite comparar as três colecções de dados, de modo quase "perfeito". A transformação efectuada não só estabilizou a dispersão, como tornou as colecções simétricas, i.e., a transformação efectuada revelou-se excelente. As amostras revelam caudas muito longas mas, enquanto as amostras recolhidas nos textos da Maria Teresa e da Maria Velho têm as caudas inferiores e superiores com o mesmo comprimento, a amostra recolhida para a Maria Isabel apresenta uma cauda direita mais longa. De salientar ainda, o facto das amostras para Maria Isabel e para a Maria Velho apresentarem uma maior dispersão. Outro aspecto a reter é o facto da amostra para a Maria Teresa ter dois outliers inferiores.

Quanto ao comprimento de frase utilizado pelas autoras, pode-se concluir que: a Maria Isabel é a que utiliza frases mais longas, a Maria Velho e a Maria Teresa

utilizam, em termos médios, frases sensivelmente do mesmo tamanho; no entanto, a Maria Velho é, das três, a que utiliza frases mais curtas.

Ainda para o comprimento de frase nos textos de autoria conhecida e, uma vez que a média e a mediana são bastante diferentes para as três amostras (a média é sempre superior à mediana), resolveu-se calcular o seu coeficiente de assimetria. Os dados obtidos foram os seguintes:

Autoras	Coeficiente de assimetria
Isabel	3.973
Teresa	2.544
Velho	1.841

Tabela 44: Coeficientes de assimetria para as amostras recolhidas nos textos de autoria conhecida

Verifica-se que o enviesamento é positivo e relativamente grande para as três amostras; a que foi recolhida para a Maria Isabel é a que apresenta um maior enviesamento, seguida da recolhida para a Maria Teresa; sendo a amostra recolhida para a Maria Velho a que apresenta o menor enviesamento.

De seguida, apresenta-se um estudo análogo ao anterior, para o comprimento de frase utilizada nos textos de autoria desconhecida. Para estes textos considerou-se todas as frases que os constituíam.

Textos	N	Média	Mediana	Mínimo	Máximo	Desvio Padrão
1ª carta I	10	17.6	15	3	34	10.46
1ª carta II	24	28.88	16.5	3	160	33.53
1ª carta III	8	36.63	28.5	19	71	18.21
1ª Última(1)	21	29.05	21	6	104	24.22
1ª Última (2)	26	23.58	16.5	2	75	18.55
1ª Última (3)	38	25.92	18	3	85	22.92
2ª Carta Última	81	26.57	18	2	198	29.34
3ª Carta Última	12	15	9.5	5	41	11.56
Paz	25	21.08	17	4	63	15.28
Lamento	62	14.47	11	3	55	11.02
Monólogo	10	68.5	40	4	291	86.85
Escriturário	18	26.17	26.5	4	48	13.47
Tarefas	31	40.55	35	4	155	30.53

Tabela 45: Algumas medidas calculadas para as amostras recolhidas nos textos de autoria desconhecida

Analisando a tabela anterior, três textos sobressaem, pelo facto de terem um comprimento médio de frase bastante grande. São eles: “Primeira Carta III”, “Monólogo” e “Tarefas”. Tal facto foi entendido como sendo um indício de que a autora dos três textos poderia ser a mesma. Outro aspecto importante a salientar é, novamente, a diferença considerável, entre o valor da média e o valor da mediana a mediana é sempre mais pequena que a média, ou seja, mais uma vez se está perante colecções assimétricas à direita.

De salientar, também, a grande dispersão de todas as amostras, ou seja, mais uma vez, no mesmo texto, as três autoras utilizam frases muito grandes e, também, frases muito pequenas, o que pode ser confirmado pela análise dos valores do máximo e do mínimo.

Apresenta-se de seguida as caixas-com-bigodes paralelas para as amostras recolhidas nos textos de autoria desconhecida.

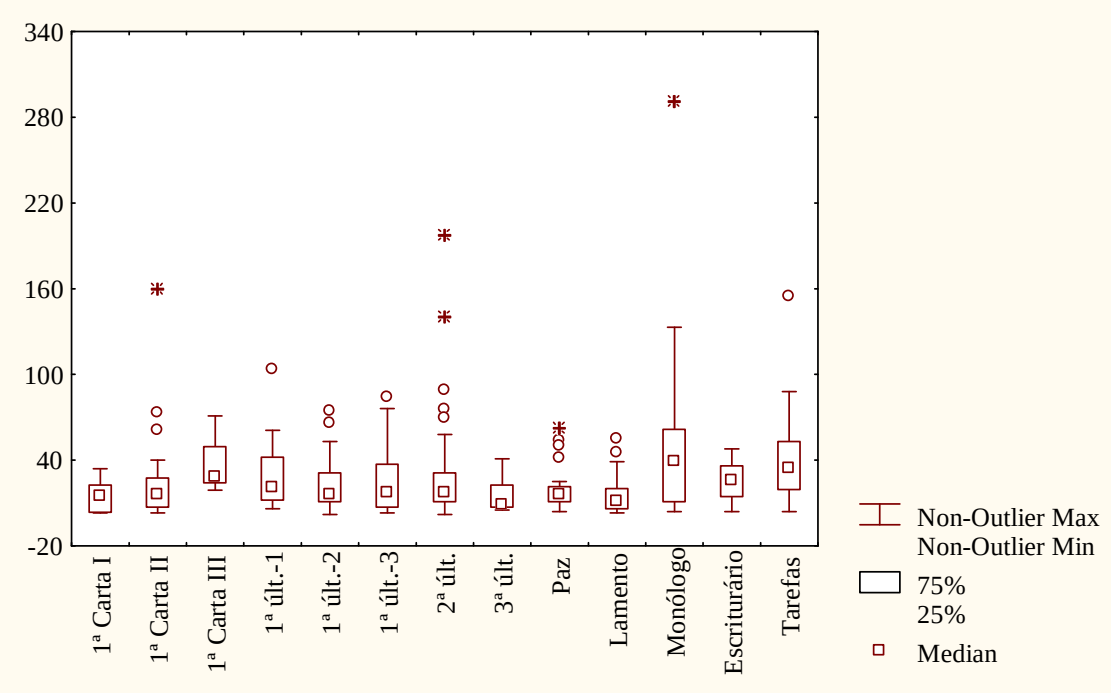

Figura 3: Caixas-com-bigodes paralelas para as amostras dos textos de autoria desconhecida

Uma vez mais, a figura anterior releva-se de difícil análise, devido à relação entre a dispersão e o nível. De salientar, no entanto, o facto de a maioria das amostras, como já tinha sido dito, revelar uma assimetria positiva e a existência de outliers. Antes de se prosseguir na análise destes dados, vai-se tentar, através de uma transformação potência, estabilizar a relação entre a dispersão e o nível.

Depois de construído o gráfico dispersão-versus-nível, concluiu-se que o declive da recta ajustada aos pontos do gráfico era b=0.761, ou seja, o expoente a utilizar na transformação potência era de p=0.239. Como 0.239 é um valor próximo de zero, utilizou-se a transformação logaritmo.

Apresentam-se, de seguida, as caixas-com-bigodes paralelas para os dados logaritmizados.

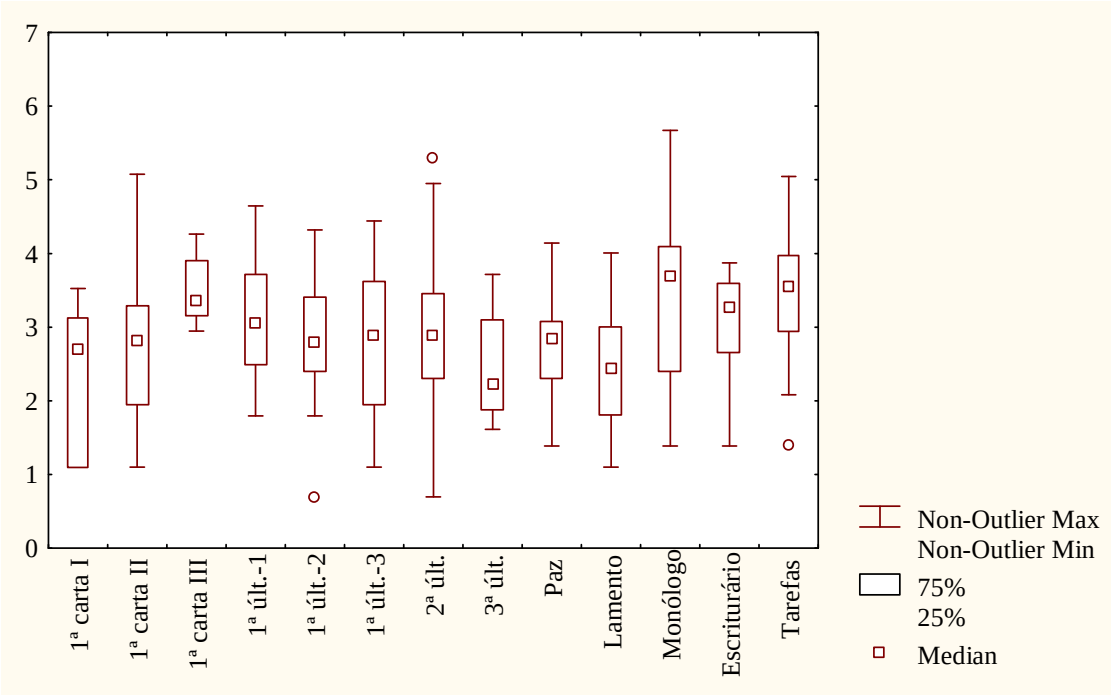

Figura 4: Caixas-com-bigodes para os dados logaritmizados

Mais uma vez, a transformação potência relevou-se um óptimo instrumento para estabilizar a dispersão; além disso, a transformação efectuada tornou algumas das colecções simétricas. Note-se o facto de a maioria dos outliers que apareciam nos dados originais, terem deixado de ser considerados como tal. Apenas três colecções apresentam outliers; são elas: as colecções recolhidas nos textos "Primeira Carta Última - Parte 2", "Tarefas" e "Segunda Carta Última". As duas primeiras colecções apresentam um outlier inferior cada uma, e a última, um outlier superior.

Antes de se compararem os textos de autoria desconhecida com os textos de autoria conhecida, relativamente ao comprimento de frase utilizado, apresenta-se de seguida os coeficientes de assimetria para as amostras recolhidas nos textos de autoria desconhecida.

Textos	**Coeficiente de assimetria**
1ª Carta I	.159773
1ª Carta II	3.148187
1ª Carta III	1.072460
1ª Carta Última - Parte1	1.660080
1ª Carta Última - Parte2	1.445862
1ª Carta Última - Parte2	1.310270
2ª Carta Última	3.509651
3ª Carta Última	1.255831
Paz	1.573244
Lamento	1.639057
Monólogo	2.233139
Escriturário	.033599
Tarefas	1.907845

Tabela 46: Coeficiente de assimetria para os dados dos textos de autoria desconhecida

Como já tinha sido dito, e a tabela anterior confirma, as amostras dos dados originais, têm uma simetria positiva forte; as amostras dos textos "Primeira Carta II", "Segunda Carta Última", "Monólogo" e "Tarefas" apresentam uma forte assimetria positiva. As restantes amostras, apesar de terem um coeficiente de assimetria mais pequeno, têm, também, uma assimetria positiva bem vincada.

Apresenta-se de seguida, as caixas-com-bigodes paralelas que permitem comparar os textos de autoria desconhecida com os de autoria conhecida, relativamente ao comprimento de frase utilizado pelas autoras. Para uma melhor visualização, dividiram-se os textos em dois conjuntos.

Conjunto das cartas

Para as amostras deste conjunto, e numa primeira fase, construiu-se o gráfico de dispersão-versus-nível, para assim se determinar uma transformação potência que estabilizasse a dispersão. Encontrou-se o valor b=0.89, como declive da recta ajustada aos pontos do gráfico, ou seja, p=0.11 seria o valor do expoente a utilizar na transformação potência, pelo que se utilizou, a transformação logaritmo. As caixas-com-bigodes para os dados logaritmizados apresentam-se a seguir.

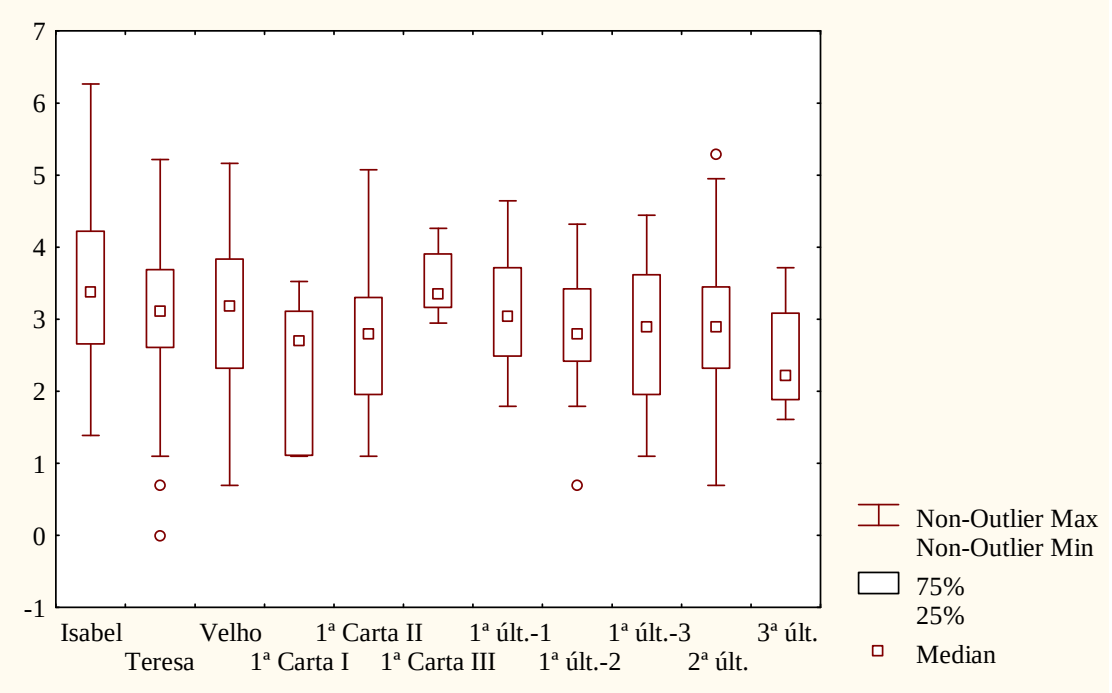

Figura 5: Caixas-com-bigodes para os dados logaritmizados

Da figura anterior, o primeiro aspecto a salientar é a existência de outliers para as colecções recolhidas para a Maria Teresa, para a "Primeira Carta Última – Parte 2" e para a "Segunda Carta Última", tendo as duas primeiras outliers inferiores e a última um outlier superior.

Se a partir da figura anterior, se se pretender inferir sobre a autoria dos textos de autoria desconhecida, chega-se às seguintes conclusões:

Texto	Tendência
1ª Carta I	*Velho*
1ª Carta II	Teresa
1ª Carta III	Isabel
1ª Carta Última - Parte 1	Velho
1ª Carta Última - Parte 2	Velho
1ª Carta Última - Parte 3	Velho
2ª Carta Última	Teresa
3ª Carta Última	Velho

Tabela 47: Uma atribuição de autoria

Note-se que, atendendo ao facto de o valor mediano dos textos da Maria Teresa e da Maria Velho serem muito próximos, todos os textos que se atribuiu à Maria Teresa podem ser da Maria Velho e vice-versa.

Numa tentativa de ultrapassar a situação descrita anteriormente, tentou-se comparar os textos considerando o coeficiente de assimetria das amostras recolhidas. Os resultados a que chegou foram os seguintes:

Texto	Tendência
1ª Carta I	Velho
1ª Carta II	Teresa
1ª Carta III	Velho
1ª Carta Última-Parte 1	Velho
1ª Carta Última-Parte 2	Velho
1ª Carta Última-Parte 3	Velho
2ª Carta Última	Isabel
3ª Carta Última	Velho

Tabela 48: Uma atribuição de autoria

Este critério entra em conflito com o utilizado anteriormente em dois textos. Na "Primeira Carta III", que com base no critério anterior se atribuiu à Maria Teresa, e agora se aproxima mais dos valores da Maria Velho, e o texto "Segunda Carta Última", que parecia ser, também, um texto da Maria Teresa, e agora, tem valores próximos dos da Maria Isabel. Para os restantes textos as conclusões são coincidentes.

Conjunto dos restantes textos

Mais uma vez se começou por traçar o gráfico de dispersão-versus-nível, para encontrar a potência da transformação a aplicar aos dados, com vista a estabilizar a dispersão. Encontrou-se o valor 1.166, como declive da recta ajustada aos dados, i. e., mais uma vez, é recomendada transformação logaritmo.

De seguida, apresentam-se, as caixas-com-bigodes paralelas para os dados transformados.

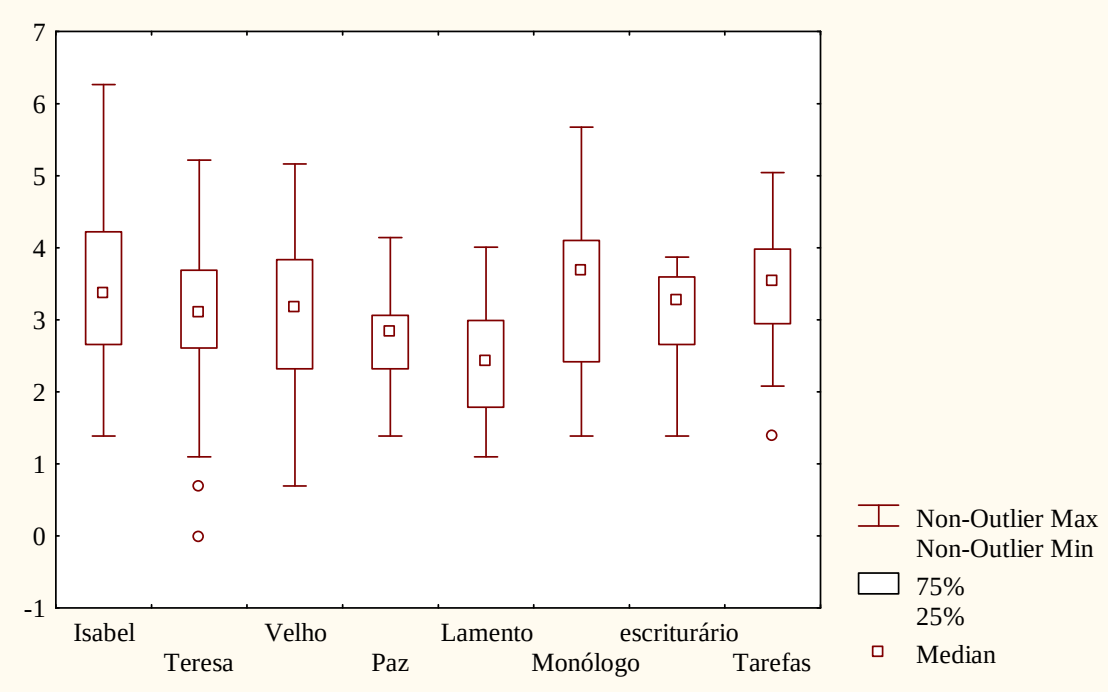

Figura 6:Caixas-com-bigodes paralelas para os dados logaritmizados

Mais uma vez, a transformação logaritmo também se mostrou eficiente para com os dados, pois, esta transformação, não só conseguiu estabilizar a dispersão, mas também tornar as colecções simétricas, excepção feita para as colecções dos textos "Paz" e "Monólogo", que apresentam uma assimetria negativa.

Recorrendo ao critério de comparação dos valores medianos do comprimento de frase, chegou-se às seguintes conclusões:

Texto	**Tendência**
Paz	Velho
Lamento	Velho
Monólogo	**Isabel**
Escriturário	Isabel
Tarefas	**Isabel**

Tabela 49: Uma atribuição de autoria

Os textos atribuídos à Maria Velho podem, em termos de valor mediano do comprimento de frases, ser atribuídos à Maria Teresa, uma vez que estas autoras usam um comprimento mediano de frase sensivelmente igual.

Apresentam-se, de seguida, os resultados das comparações efectuadas entre os coeficientes de assimetria:

Textos	***Tendência***
Paz	Velho
Lamento	Velho
Monólogo	Teresa
Escriturário	-
Tarefas	Velho

Tabela 50: Uma atribuição de autoria

Para dois dos anteriores textos, este critério entra em conflito com o critério que compara o comprimento mediano dos textos conhecidos e desconhecidos, e é omisso relativamente a um terceiro. O coeficiente de assimetria da amostra retirada do texto "Escriturário" é 0.036, ou seja, a amostra pode ser considerada simétrica, o que não se verifica para nenhuma das autoras; optou-se, então, por não fazer nenhuma

atribuição de autoria para este texto. Os que “mudaram de autor” foram o “Monólogo” e as “Tarefas”, que pelo critério anterior pertenciam à Maria Isabel, e agora, passaram a ser atribuídos à Maria Teresa e à Maria Velho, respectivamente.

Em conclusão, utilizando os dois critérios anteriores, fizeram-se as seguintes atribuições de autoria:

Textos	**Tendência (mediana)**	**Tendência (assimetria)**
1ª Carta I	Velho	Velho
1ª Carta II	Teresa	Teresa
1ª Carta III	**Isabel**	**Velho**
1ª Carta Última - Parte 1	Velho	Velho
1ª Carta Última - Parte 2	Velho	Velho
1ª Carta Última - Parte 3	Velho	Velho
2ª Carta Última	**Teresa**	**Isabel**
3ª Carta Última	Velho	Velho
Paz	Velho	Velho
Lamento	Velho	Velho
Monólogo	**Isabel**	**Teresa**
Escriturário	Isabel	-
Tarefas	**Isabel**	**Velho**

Tabela 51: Uma atribuição de autoria — resumo

Ou seja, exceptuando os textos “Primeira Carta III”, “Segunda Carta Última”, “Monólogo” e "Tarefas", os dois critérios utilizados caminham na mesma direcção, i.e., a autoria dos textos é atribuída sempre à mesma autora.

O comprimento de frase apresenta-se como sendo uma variável distintiva das três autoras, ou, pelo menos, entre uma delas e as outras duas. Pode dizer-se, como já foi referido, que a Maria Isabel utiliza as frases mais longas, sendo por isso “relativamente fácil” distinguir os seus textos; já a Maria Teresa e a Maria Velho utilizam frases com, sensivelmente, o mesmo tamanho, o que torna a variável pouco eficiente para distinguir estas duas autoras entre si.

Das variáveis até aqui estudadas, esta parece ser uma das que melhor consegue discriminar as autoras; é, por isso, uma variável a ter em atenção em futuros trabalhos.

3.4. Comprimento de parágrafo

Outra variável que pareceu ser de investigar, foi o tamanho dos parágrafos utilizados pelas autoras. A leitura das obras, não só despertou a atenção para o comprimento de frase utilizado, mas também, para o comprimento dos parágrafos. Aquando da leitura das obras, o tamanho, por vezes extraordinário, dos parágrafos, chamou a atenção, e, como tal, decidiu-se investigar a variável comprimento de parágrafo.

Mais uma vez, recorreu-se aos blocos de texto de autoria conhecida para definir um comprimento padrão de parágrafo para cada autora. Tal como foi referido na subsecção anterior, o início e o fim dos blocos de texto, de autoria conhecida, nem sempre coincidiu com o início e o fim de frases, e muito menos de parágrafos. Como se decidiu estudar o comprimento de todos os parágrafos que constituíam os blocos conhecidos, duas alternativas se colocaram para contornar o problema. Ou se eliminava o primeiro e o último parágrafo do bloco, sempre que o início e o fim de bloco não coincidia com o início e o fim do parágrafo, ou se completava o referido bloco. Optou-se pela segunda alternativa, pois a eliminação do primeiro e do último parágrafo, tinha como consequência, por vezes, uma redução drástica do corpus de estudo. Assim, sempre que o início e/ou o fim do bloco não coincidisse com o início e/ou o fim de parágrafo, o bloco era aumentado de modo a começar e a acabar num parágrafo. Deste modo, e nesta subsecção não se trabalhou com blocos de textos com mil e quinhentas palavras, mas, com blocos de texto de comprimento variável, mas sempre, igual ou superior a mil e quinhentas palavras.

O comprimento de parágrafo foi, como é óbvio, avaliado pelo número de palavras que o constituíam. Para cada um dos blocos de texto conhecido, disponíveis para cada autora, contou-se o número de parágrafos que o constituíam, e o número de palavras que formavam cada um dos parágrafos.

A tabela seguinte apresenta o número total de parágrafos encontrados nos blocos de autoria conhecida, bem como algumas medidas calculadas para os dados

recolhidos. Apresenta-se, assim, a média, a mediana, o máximo, o mínimo e o desvio-padrão do número de palavras por parágrafo.

Autora	N	Média	Mediana	Mínimo	Máximo	Desvio Padrão
Isabel	124	168.45	90.500	2.00	1231.0	210.02
Teresa	78	212.95	167.00	1.00	675.0	185.56
Velho	106	172.69	105.50	1.00	1535.0	211.82

Tabela 52 : Algumas medidas para o tamanho de parágrafo utilizado pelas autoras

A primeira conclusão a retirar da tabela anterior, é o facto da diferença entre as médias e as medianas ser significativa, i.e., as amostras apresentam caudas direitas muito pesadas. Pode ainda concluir-se que a Maria Teresa é a que utiliza, em termos médios, parágrafos maiores, sendo, também, a autora para a qual a amostra recolhida apresenta uma menor dispersão. As outras duas autoras utilizam um tamanho de parágrafo idêntico, e, em média, mais pequeno do que a Maria Teresa; no entanto, e por vezes, utilizam parágrafos extremamente longos. Note-se que estas duas autoras utilizaram parágrafos com mais de mil palavras; no caso da Maria Velho, um dos parágrafos tem mais de mil e quinhentas palavras, ou seja, um dos blocos escolhidos é constituído apenas por um parágrafo. O facto de a Maria Isabel e da Maria Velho, utilizarem parágrafos pequenos, mas também parágrafos longos, explica o valor do desvio-padrão para as duas autoras, que é bastante grande.

De seguida, apresentam-se, as caixas-com-bigodes paralelas para as três amostras recolhidas.

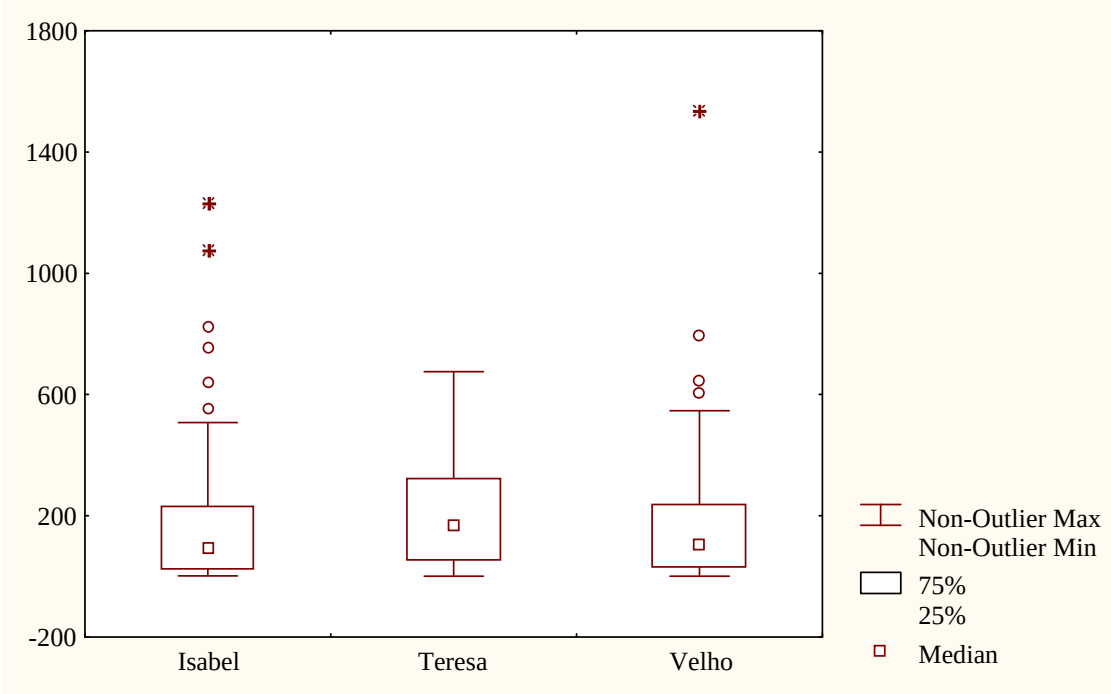

Figura 7: Caixas-com-bigodes para as amostras recolhidas nos textos conhecidos

Da análise da figura anterior, concluiu-se, que as amostras recolhidas para as três autoras, apresentam caudas direitas longas e caudas esquerdas curtas, sendo a amostra recolhida para a Maria Teresa a que apresenta caudas mais compridas. Saliente-se o facto, de as amostras das três autoras apresentarem assimetria positiva; no entanto, e mais uma vez, a amostra recolhida para a Teresa Horta distingue-se das outras duas por ser a que apresenta uma assimetria mais pequena. É aliás, uma amostra quase simétrica. O aspecto mais evidente do gráfico anterior, reside no facto das amostras para a Maria Isabel e para a Maria Velho apresentarem outliers; a amostra da Maria Isabel contém seis outliers, e a amostra da Maria Velho contém quatro.

Para tentar eliminar a relação entre a dispersão e o nível existente nos dados, aplicou-se a estes uma transformação potência. Para tal começou-se por traçar o gráfico dispersão-versus-nível, para identificar a transformação a efectuar. O declive encontrado para a recta ajustada aos pontos do gráfico, foi de 0.413, pelo que p=(1-0.413)=0.587 seria o valor aproximado do expoente da transformação a utilizar, para

estabilizar a dispersão. Como p=0.587 é um valor próximo de 0.5 efectuou-se a transformação-raiz quadrada. Apresenta-se, de seguida, as caixas-com-bigodes paralelas para a raiz-quadrada dos dados.

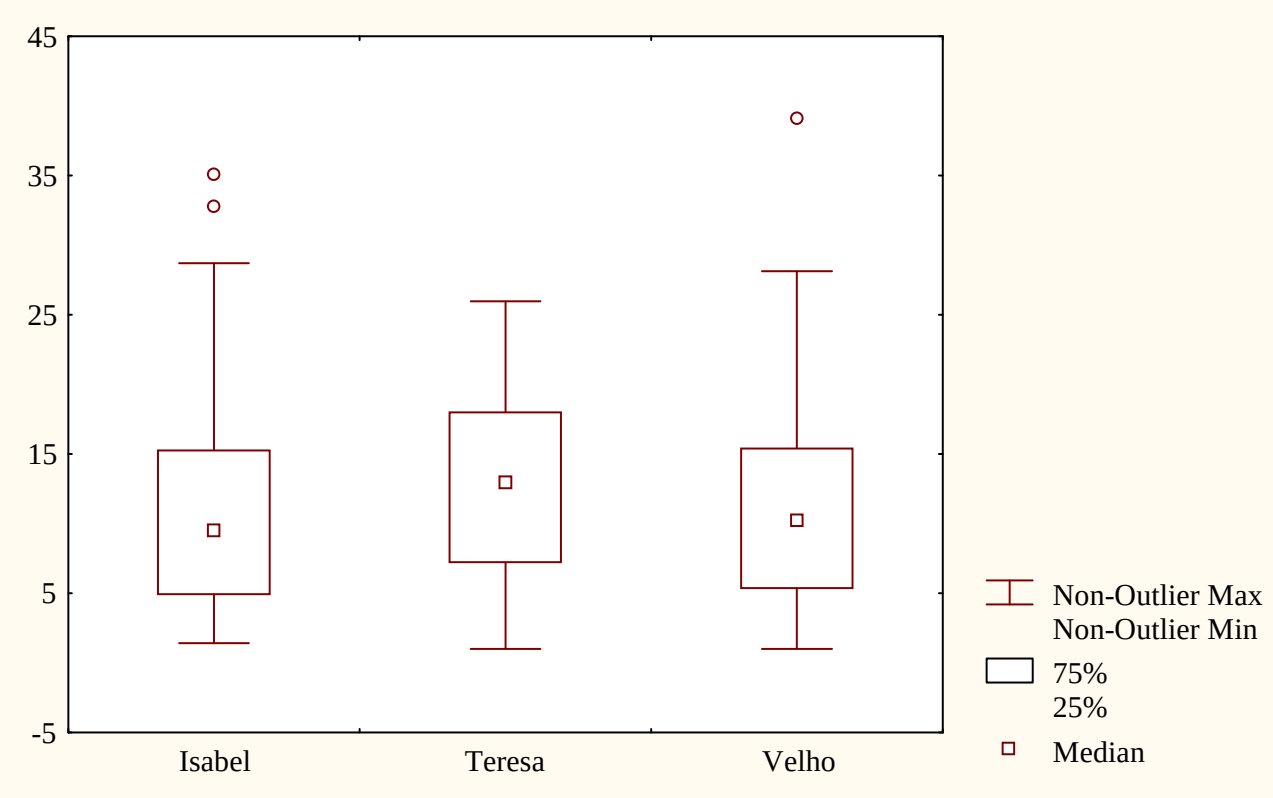

Figura 8: Caixas-com-bigodes paralelas para a raiz-quadrada dos dados

Da figura anterior, concluiu-se que a relação entre a dispersão e o nível foi eliminada; além disso, a assimetria das colecções foi também eliminada, tendo-se agora três colecções simétricas. Como já foi referido, as amostras recolhidas para a Maria Isabel e para a Maria Velho, apresentam caudas direitas longas e caudas esquerdas curtas; a amostra recolhida para a Maria Teresa tem comprimentos de cauda sensivelmente iguais. Outro aspecto importante a salientar, foi a eliminação da

maioria dos outliers das amostras da Maria Isabel e da Maria Velho; estas amostras apresentam, agora, apenas dois e um outlier, respectivamente.

Como a diferença entre o valor da média e da mediana é significativo, calcularam-se os coeficientes de assimetria para as três amostras. Os resultados obtidos apresentam-se na tabela seguinte:

Autora	Coeficiente de Assimetria
Isabel	2.405817
Teresa	.845259
Velho	3.144247

Tabela 53 : Coeficientes de assimetria para as amostras recolhidas nos textos de autoria conhecida

Da análise da tabela anterior, conclui-se, como seria de esperar, que as amostras para a Maria Isabel e para Maria Velho tem um enviesamento à direita bastante acentuado, enquanto para a Maria Teresa esse enviesamento, embora inegável, é bastante inferior. Tal facto, vem confirmar que a Maria Teresa não utiliza parágrafos com um comprimento tão “extremo” como, por vezes, as colegas.

Depois de efectuado o estudo anterior, houve que voltar aos textos de autoria desconhecida e recolher, para cada um, o número de parágrafos que o constituía, assim como, o número de palavras que constituía cada parágrafo.

Apresentam-se, de seguida, algumas medidas de localização e dispersão calculadas para as amostras recolhidas.

Textos	N	Média	Mediana	Mínimo	Máximo	Desvio Padrão
1ª carta I	3	58.67	55.00	35.00	86.00	25.71
1ª carta II	12	42.17	36.00	13.00	90.00	25.14
1ª carta III	3	99.67	100.00	75.00	124.00	24.50
1ª carta última-1	5	116.80	110.00	46.00	177.00	48.42
1ª carta última-2	6	101.00	93.50	15.00	201.00	81.29
1ª carta última-3	8	119.50	105.50	34.00	340.00	95.93
2ª carta última	11	195.09	186.00	67.00	369.00	107.01
3ª carta última	9	19.78	20.00	5.00	40.00	11.98
A Paz	14	37.71	26.00	4.00	113.00	31.13
Lamento	21	42.48	24.00	3.00	208.00	45.95
Monólogo	1	684.00	-	684.00	684.00	-
Escriturário	12	39.25	33.50	12.00	69.00	20.39
Tarefas	12	112.92	69.500	20.00	529.00	137.67

Tabela 54: Algumas medidas para o tamanho de parágrafo utilizado pelas autoras nos textos de autoria desconhecida

Analisando a tabela anterior, a primeira conclusão a retirar é que, de um modo geral, os textos que são constituídos por um maior número de parágrafos, apresentam, em média, um tamanho de parágrafo mais pequeno, excepção feita aos textos “Segunda Carta Última” e “Tarefas”. Os textos que apresentam parágrafos maiores são: “Primeira Carta III”, as “Primeiras Cartas Últimas”, a “Segunda Carta Última”, e os textos “Monólogo” e “Tarefas” (note-se que os textos referidos têm menos de dez parágrafos, excepto os textos “Segunda Carta Última” e “Tarefas”, com onze e doze parágrafos, respectivamente). Saliente-se, ainda, o facto de os textos “Segunda Carta Última” e “Tarefas” apresentarem a maior dispersão e da diferença entre a média e a mediana ser acentuada para estas amostras.

Atendendo ao facto, de os textos “Monólogo”, “Primeira Carta I” e “Primeira Carta III” serem constituídos por um número reduzido de parágrafos, vai-se eliminá-los da análise que se segue.

Apresenta-se, de seguida, as caixas-com-bigodes para as amostras anteriores.

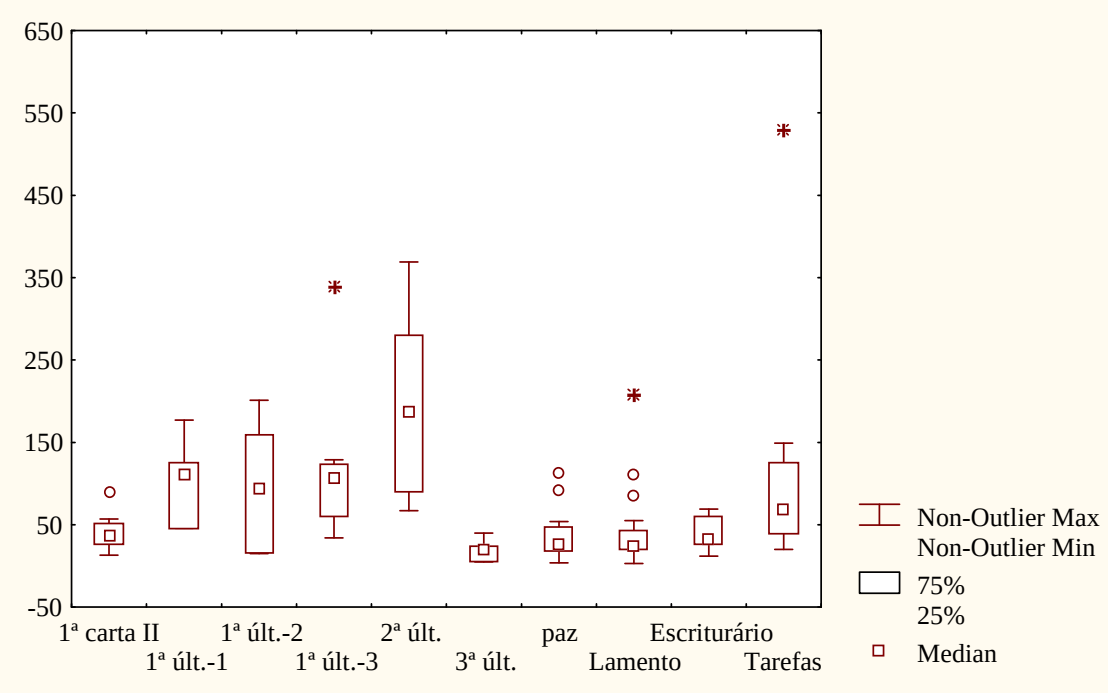

Figura 9: Caixas-com-bigodes paralelas para as amostras recolhidas nos textos de autoria desconhecida

A análise da figura anterior é difícil, devido a relação entre a dispersão e o nível. Antes de se tentar estabilizar a relação anterior através de uma transformação potência, convém referir alguns aspectos importantes da figura anterior.

Assim, existem cinco textos cujas amostras apresentam outliers; são eles: "Primeira Carta II", "Primeira Carta Última – Parte 3", "Paz", "Lamento" e "Tarefas". As amostras têm, de um modo geral, assimetria positiva, com excepção das amostras dos textos "Primeira Carta Última – Parte 1" e "Terceira Carta Última", que apresentam assimetria negativa. A amostra recolhida no texto "Segunda Carta Última", destaca-se de todas as outras, por ser simétrica e apresentar cauda direita longa.

Efectuou-se de seguida, o processo de diagnóstico para obtenção da potência da transformação que estabilizasse a dispersão. O gráfico de dispersão-versus-nível obtido deu, como declive para a recta ajustada, o valor 0.975, donde o valor

aproximado do expoente a utilizar na transformação-potência seria 0.025; ou seja, efectuou-se a transformação logaritmo dos dados.

Apresenta-se de seguida as caixas-com-bigodes paralelas para os dados logaritmizados.

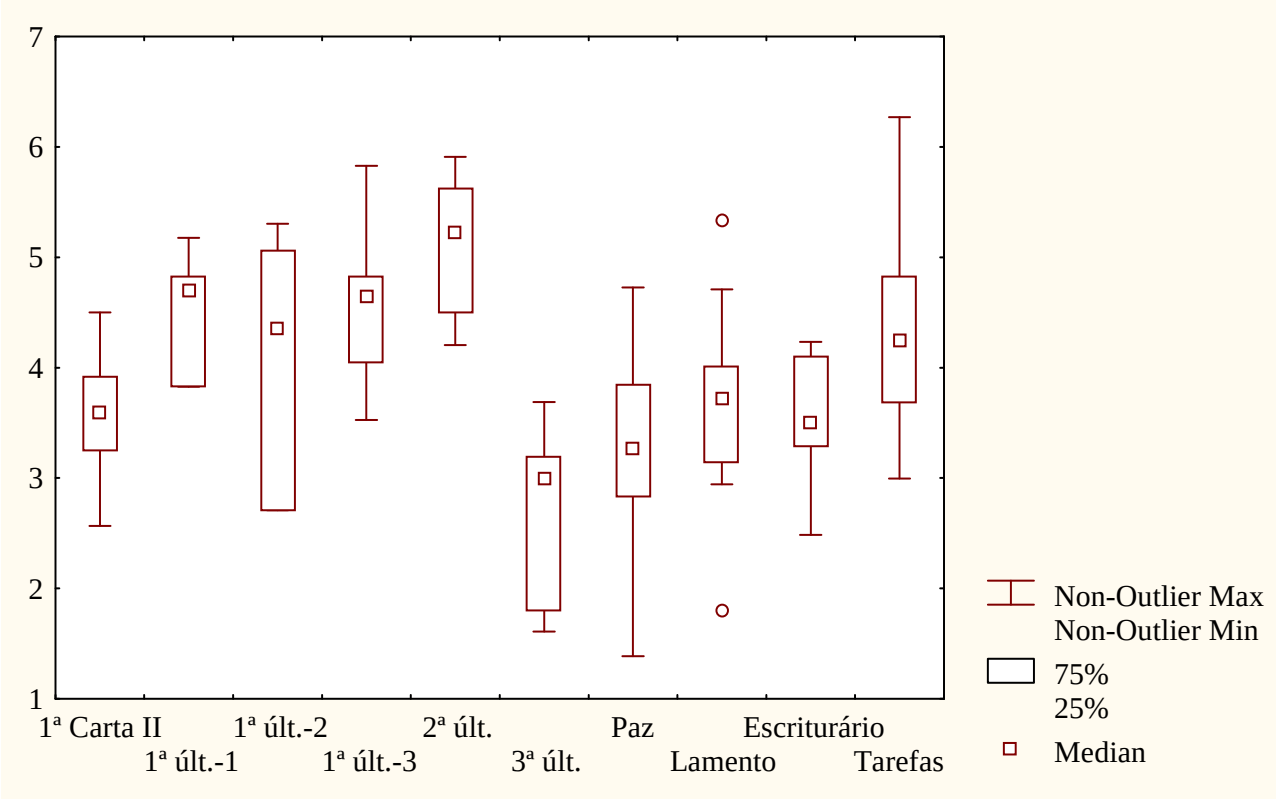

Figura 10: Caixas-com-bigodes paralelas para os dados logaritmizados

A análise da figura anterior permite concluir que a relação entre a dispersão e o nível foi estabilizada, mas não se conseguiu, com a mesma transformação, resolver o problema da assimetria. A maioria das colecções apresenta agora, assimetria negativa, com excepção da colecção recolhida no texto "Escriturário", que é assimétrica positiva, e das colecções dos textos "Primeira carta II", "Paz" e "Tarefas" que são simétricas. Um facto importante a referir, é o aparecimento de um

outlier inferior na amostra do texto “Lamento”, que na amostra original, apenas apresentava um outlier superior; saliente-se, ainda, o facto de a transformação efectuada ter removido os outliers exibidos pelas amostras dos textos: “Primeira Carta Última – Parte 3” e “Tarefas”.

Antes de se compararem os textos de autoria conhecida com os textos de autoria desconhecida, no que diz respeito ao tamanho dos parágrafos, vai-se, primeiro, calcular o coeficiente de assimetria, para as amostras originais recolhidas nos textos de autoria desconhecida.

Textos	**Coeficiente de assimetria**
1ª Carta II	1.20
1ª Carta Última-parte1	-0.45
1ª Carta Última-parte2	0.15
1ª Carta Última-parte3	2.06
2ª Carta Última	0.32
3ª Carta Última	0.39
A Paz	1.45
Lamento	2.67
Escriturário	0.39
Tarefas	2.91

Tabela 55: Coeficiente de assimetria para as amostras recolhidas nos textos de autoria desconhecida

Analisando a tabela anterior conclui-se, que a maioria das amostras recolhidas apresenta um enviesamento positivo; no entanto, o enviesamento apresentado pelas amostras, não é, de um modo geral, muito acentuado, excepção feita para as amostras recolhidas nos textos “Primeira Carta II”, “Primeira Carta Última – Parte 3”, “Paz”, “Lamento” e “Tarefas”, que apresentam uma simetria à direita forte.

Saliente-se o facto, da amostra do texto “Primeira Carta Última - Parte 1” apresentar uma assimetria negativa.

Numa tentativa de “descobrir“ a autoria dos textos “desconhecidos”, apresentam-se, de seguida, as caixas-com-bigodes paralelas para os textos de autoria conhecida e para os textos de autoria desconhecida. Para melhor se visualizarem as figuras, vão-se separar os textos de autoria desconhecida em dois conjuntos; um, constituído pelas primeiras e últimas cartas, e outro, pelos restantes textos. Mais uma vez, e

pelas razões atrás apontadas, não foram considerados os textos “Primeira Carta I e III” e o “Monólogo”.

Conjunto das cartas

Traçado o gráfico de dispersão-versus-nível para os dados, chegou-se a 1.195 para o valor do declive da recta ajustada; ou seja, 0.195 é o valor aproximado do expoente da transformação-potência. Assim, efectuou-se a transformação logaritmo nos dados.

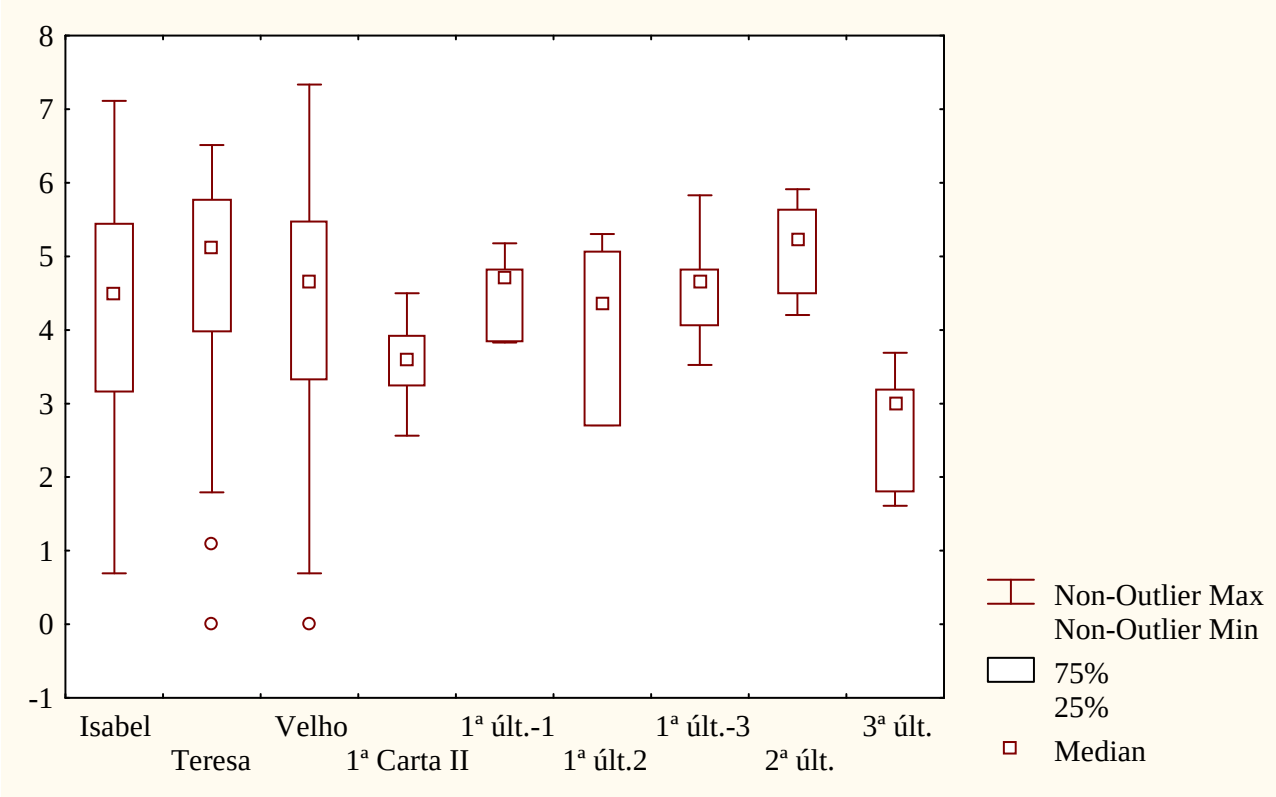

Figura 11: Caixas-com-bigodes paralelas para os dados logaritmizados

Antes de se prosseguir com a análise, convém referir o aparecimento de dois outliers inferiores para a amostra recolhida para a Maria Teresa, e um outlier inferior, na amostra recolhida para a Maria Velho.

Analisando a figura anterior, conclui-se que, em termos de valor mediano, talvez a Isabel Barreno tenha escrito os textos “Primeira Carta II”, “Primeira Carta Última - Parte 2” e “Terceira Carta Última”; a Maria Velho, talvez seja a autora dos textos “Primeira Carta Última - Partes 1 e 3”; sendo a Teresa Horta, talvez, a autora da “Segunda Carta Última”.

Da análise do gráfico anterior, concluiu-se que a Maria Velho talvez tenha escrito os textos “Primeira Carta Última - Partes 1 e 3”, e que a Maria Isabel, talvez, seja a autora da “Primeira Carta Última - Parte 2”; esta conclusão está em contradição com a assunção feita, no início deste trabalho, de que as três partes que constituem a

"Primeira Carta Última" são da mesma autora. Tal facto vem reforçar a ideia, que, talvez, uma das três autora tenha feito um trabalho de revisão dos textos. Se tal hipótese for verdadeira, a Maria Isabel ou a Maria Velho afiguram-se como as possíveis revisoras dos textos.

Conjunto dos restantes textos

Também neste caso, começou-se por determinar a transformação potência para estabilizar a relação entre a dispersão e o nível. A inclinação da recta ajustada ao gráfico dispersão-versus-nível, apresentou um declive de 1.356, ou seja, obteve-se para valor aproximado do expoente, da transformação potência, o valor 0.346. Como 0.346 está mais perto de 0.5 do que de zero, optou-se pela transformação-raiz quadrada. As caixas-com-bigodes paralelas, que a seguir se apresentam, foram construídas para a raiz quadrada dos dados.

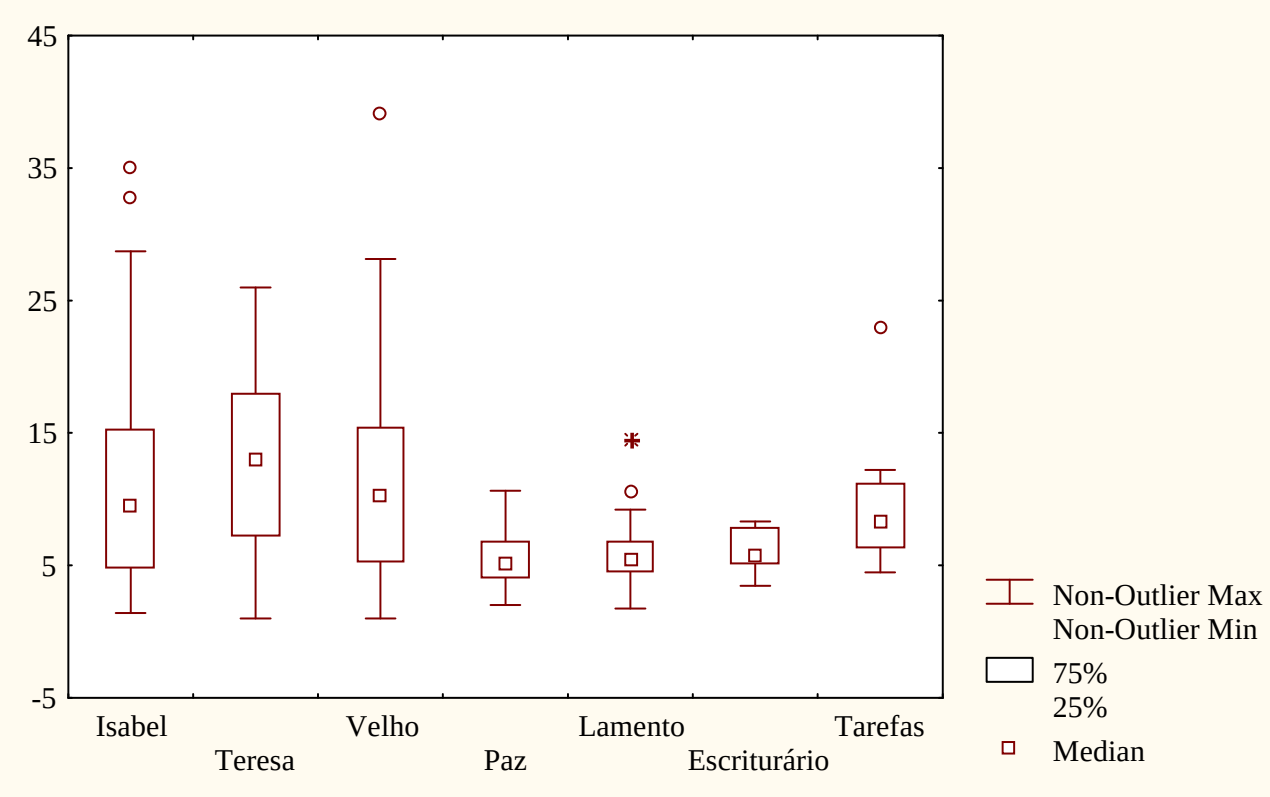

Figura 12: Caixas-com-bigodes paralelas para a raiz quadrada dos dados

Note-se, desde já, a existência de outliers superiores em quatro das amostras. A amostra recolhida nos textos da Maria Isabel apresenta, nesta escala, dois outliers superiores, assim como a amostra recolhida no texto "Lamento"; já as amostras recolhidas para Maria Velho e no texto "Tarefas", apresentam, cada uma, um outlier superior.

Efectuando uma análise análoga à efectuada para o primeiro conjunto de textos, conclui-se, que a autora de todos os textos é a Maria Isabel, pois, esta autora, é a que tem um comprimento mediano de parágrafo, mais próximo do valor apresentado pelos textos desconhecidos.

Se se utilizar o coeficiente de assimetria para tirar algumas conclusões, chega-se à conclusão de que exceptuando quatro textos, todos os outros têm o valor do coeficiente de assimetria próximo do valor do coeficiente de assimetria da Maria Teresa.

Textos	Tendência
1ª Carta II	Teresa
1ª Carta Última-parte1	-
1ª Carta Última-parte2	Teresa
1ª Carta Última-parte3	Isabel
2ª Carta Última	Teresa
3ª Carta Última	Teresa
A Paz	Teresa
Lamento	Isabel
Escriturário	Teresa
Tarefas	Velho

Tabela 56: Uma atribuição de autoria — resumo

Ou seja, a utilização do coeficiente de assimetria para fazer uma atribuição de autoria, só veio lançar mais confusão para o estudo. Se, até aqui, a Maria Teresa se apresentava como a autora menos provável da maioria dos textos, subitamente, ela aparece como uma das autoras mais prováveis.

Resumindo,

Textos	Critério da mediana	Critério da assimetria
1ª Carta II	**Isabel**	**Teresa**
1ª Carta Última-parte1	Velho	-
1ª Carta Última-parte2	**Isabel**	**Teresa**
1ª Carta Última-parte3	**Velho**	**Isabel**
2ª Carta Última	Teresa	Teresa
3ª Carta Última	**Isabel**	**Teresa**
A Paz	**Isabel**	**Teresa**
Lamento	Isabel	Isabel
Escriturário	**Isabel**	**Teresa**
Tarefas	**Isabel**	**Velho**

Tabela 57: Uma atribuição de autoria

Para o texto “Primeira Carta Última – Parte 1”, não se fez atribuição de autoria, porque a amostra recolhida para este texto apresenta um coeficiente de assimetria negativo, o que não se verifica para nenhuma das autoras.

Note-se que os dois critérios anteriores só “estão de acordo” para os textos “Segunda Carta Última” e “Lamento”, cujas autoras apontadas são a Maria Teresa e a Maria Isabel, respectivamente. Para os restantes textos tudo permanece em aberto.

Antes de se finalizar esta subsecção, saliente-se que para os três textos retirados do estudo anterior, o único para o qual se pode dizer mais alguma coisa sobre a sua autoria, é o texto “Monólogo”; quer a Maria Isabel quer a Maria Velho, apresentam parágrafos de dimensão aproximada à dimensão do único parágrafo que constitui o referido texto; pelo que a autoria deste texto talvez deva ser discutida entre as autoras referidas. Os textos “Primeira Carta I e II” são muito pequenos, resultando daí a dificuldade em dizer algo mais sobre a sua autoria.

A constatação dos factos anteriores, apenas veio lançar ainda mais confusão, sobre as conclusões até aqui retiradas e criar a necessidade de investigar mais variáveis. De salientar, que a hipótese de uma das autoras ter efectuado a revisão dos textos, é a que ganha cada vez mais consistência; e, se até aqui, a Maria Isabel e a Maria Velho se apresentavam como as possíveis revisoras, as conclusões anteriores vieram, de novo, baralhar tudo.

3. 5. Pontuação utilizada

Outra das variáveis que se decidiu estudar foi o tipo de pontuação utilizada e a sua frequência. A escolha desta variável ficou a dever-se, também, à leitura das obras em estudo, pois verificou-se que as autoras recorriam a vários sinais de pontuação, sendo alguns deles, utilizados com bastante frequência. Os textos estão “contaminados” de travessões, aspas, parêntesis, pontos e vírgulas, etc., facto que despertou a nossa atenção, e daí a decisão de investigar esta variável.

Decidiu-se não estudar, em particular, dois sinais de pontuação — a vírgula e o ponto final. O facto destes sinais poderem ser considerados como sinais padrão da língua portuguesa esteve na base da decisão tomada. O ponto final e a vírgula são dois dos sinais de pontuação mais utilizados na língua portuguesa e como tal, a sua

frequência de utilização, não pareceu ser uma variável importante para distinguir as autoras. Além disso, quer a vírgula quer o ponto final, têm que ver directamente com o tipo e com o comprimento de frase; a segunda característica já foi estudada neste trabalho, e a primeira vai ser abordada mais à frente.

Assim, para cada autora e para cada bloco de texto conhecido, foi verificar-se que outros sinais de pontuação, para além do ponto final e da vírgula, eram utilizados e qual a sua frequência de utilização em cada bloco.

Depois de obtidas as frequências absolutas de cada sinal de pontuação em cada um dos doze blocos (onze para a Maria Teresa), converteu-se essa frequência no número de sinais, desse tipo, esperado, num bloco com mil palavras. Ou seja, de um modo análogo ao feito com a frequência das palavras, calculou-se uma permilagem para os sinais de pontuação. As razões para tal são em tudo idênticas, às apontadas para o caso da frequência das palavras.

Depois deste cálculo, determinou-se a média, a mediana e o desvio-padrão para as três amostras, apresentando–se os resultados obtidos para cada autora, nas tabelas seguintes.

	« »	-	?	!	“ “	( )	;	:
Média	12.22	12.83	5.33	0.17	0.61	0.11	3.67	0.78
Mediana	14.67	14.67	5	0	0	0	4	0.67
Desvio padrão	6.43	3.97	3.73	0.30	1.54	0.38	1.55	1.09

Tabela 58: Algumas medidas calculadas para a amostra da Maria Isabel

Para a Maria Isabel, encontraram-se oito sinais de pontuação distintos. Os mais utilizados são as aspas, o travessão e o ponto de interrogação; os menos utilizados são o ponto de exclamação, as pelicas, os parêntesis e os dois pontos. As aspas é o sinal que apresenta maior variabilidade com um desvio-padrão de 6.43, o que significa que esta autora, num dado bloco, tanto pode utilizar as aspas frequentemente como raramente.

	-	:	;	...	?	( )	« »	!
Média	20.93	7.82	5.52	1.39	2.18	1.70	3.27	0.12
Mediana	19.67	6.67	4.67	1.33	2	1.33	3.33	0
Desvio padrão	4.92	3.21	1.98	1.13	2.05	2.14	2.97	0.40

Tabela 59: Algumas medidas calculadas para a amostra da Maria Teresa

Também nos textos da Maria Teresa se encontraram oito sinais de pontuação diferentes. Esta autora utiliza com alguma frequência o travessão, os dois pontos, o ponto e vírgula e as aspas, utilizando menos o ponto de exclamação e os parêntesis. De notar que a variabilidade associada ao uso do travessão, dos dois pontos e do ponto e vírgula é elevada, especialmente, para o travessão.

	-	‘ ‘	?	!	...	;	:
Média	14.61	3.55	4.67	0.28	0.17	0.22	0.39
Mediana	12.33	4	4	0	0	0	0
Desvio padrão	7.43	2.52	5.06	0.45	0.30	0.33	0.60

Tabela 60: Algumas medidas calculadas para a amostra da Maria Velho

Para a Maria Velho encontraram-se sete sinais diferentes, sendo o mais utilizado o travessão, que apresenta, também, a maior variabilidade; os menos utilizados são o ponto de exclamação, as reticências, o ponto e vírgula e os dois pontos.

Das tabelas anteriores conclui-se que a pontuação utilizada nos textos, talvez seja útil para ajudar a distinguir as autoras, não tanto pelo tipo de pontuação utilizada — tirando um ou outro sinal as autoras utilizam, sensivelmente, os mesmos — mas pela frequência com que utilizam cada sinal.

Como para os textos de autoria conhecida, o travessão, o ponto de interrogação e as aspas, apresentam uma variabilidade muito grande, resolveu-se eliminá-los do estudo que se segue. De facto, sinais com grande variabilidade de utilização não ajudam a identificar as autoras, à semelhança do que se passava com a variabilidade

das palavras; recorde-se que palavras com grande variabilidade não são adequadas para identificar as autoras.

Utilizando uma metodologia análoga à usada com a frequência das palavras seleccionadas, foi-se comparar a média de utilização de cada sinal de pontuação com o valor obtido nos textos de autoria desconhecida, depois de calculada a permilagem.

Como medida de tendência central, utilizou-se a média e não a mediana, pois a diferença entre estes dois valores não é considerável.

Os resultados obtidos foram os seguintes:

Sinal de pontuação	**Texto desconhecido**	**Isabel**	**Teresa**	**Velho**	**Tendência**
:	5.68	0.78	7.82	0.39	Teresa
;	11.36	3.67	5.52	0.22	Teresa
					Teresa

Tabela 61: Resultados para a Primeira Carta I

Sinal de pontuação	**Texto desconhecido**	**Isabel**	**Teresa**	**Velho**	**Tendência**
:	15.81	0.78	7.82	0.39	Teresa
;	1.98	3.67	5.52	0.22	Isabel
...	1.98	-	1.39	0.17	Teresa
(.)	1.98	0.11	1.70	-	Teresa
					Teresa

Tabela 62: Resultados para a Primeira Carta II

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
:	3.34	0.78	7.82	0.39	Isabel
(.)	3.34	0.11	1.70	-	Teresa
					Isabel/Teresa

Tabela 63: Resultados para a Primeira Carta III

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
(.)	3.31	0.11	1.70	-	Teresa
!	1.65	0.17	0.12	0.28	Velho
					Isabel/Velho

Tabela 64: Resultados para a Primeira Última – Parte 1

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
(.)	6.54	0.11	1.70	-	**Teresa**

Tabela 65: Resultados para a Primeira Carta Última – Parte 2

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
:	3.14	0.78	7.82	0.39	Isabel
;	1.05	3.67	5.52	0.22	Velho
...	2.09	-	1.39	0.17	Teresa
(.)	6.28	0.11	1.70	-	Teresa
					Teresa

Tabela 66: Resultados para a Primeira Carta Última – Parte 3

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
:	3.14	0.78	7.82	0.39	Isabel
;	1.05	3.67	5.52	0.22	Velho
...	2.09	-	1.39	0.17	Teresa
(.)	6.28	0.11	1.70	-	Teresa
					Teresa

Tabela 67: Resultados para a Segunda Carta Última

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
:	11.24	0.78	7.82	0.39	Teresa
;	5.62	3.67	5.52	0.22	Teresa
...	11.24	-	1.39	0.17	Teresa
(.)	28.09	0.11	1.70	-	Teresa
!	11.24	0.17	0.12	0.28	Velho
					Teresa

Tabela 68: Resultados para a Terceira Carta Última

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
:	7.58	0.78	7.82	0.39	Teresa
;	9.47	3.67	5.52	0.22	Teresa
...	1.89	-	1.39	0.17	Teresa
(.)	1.89	0.11	1.70	-	Teresa
					Teresa

Tabela 69: Resultados para a Paz

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
:	7.85	0.78	7.82	0.39	Teresa
;	2.24	3.67	5.52	0.22	Isabel
...	10.09	-	1.39	0.17	Teresa
(.)	4.48	0.11	1.70	-	Teresa
!	4.48	0.17	0.12	0.28	Velho
					Teresa

Tabela 70: Resultados para a Lamento

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
:	7.31	0.78	7.82	0.39	Teresa
...	8.77	-	1.39	0.17	Teresa
!	2.92	0.17	0.12	0.28	Velho
					Teresa

Tabela 71: Resultados para a Monólogo

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
:	2.11	0.78	7.82	0.39	Isabel
;	4.21	3.67	5.52	0.22	Isabel
...	2.11	-	1.39	0.17	Teresa
!	8.42	0.17	0.12	0.28	Velho
					Isabel

Tabela 72: Resultados para a Escriturário

Sinal de pontuação	Texto desconhecido	Isabel	Teresa	Velho	Tendência
:	5.89	0.78	7.82	0.39	Teresa
;	2.21	3.67	5.52	0.22	Isabel
...	0.74	-	1.39	0.17	Velho
(.)	1.47	0.11	1.70	-	Teresa
					Teresa

Tabela 73: Resultados para a Tarefas

Mais uma vez os “dados foram baralhados”. Em dez dos treze textos de autoria desconhecida, os valores encontrados indiciam como autora a Maria Teresa, facto que parece ser pouco provável. Dos restantes três textos, um (“Escriturário”) foi atribuído à Maria Isabel, enquanto, que a autoria dos textos “Primeira Carta I” e “Primeira Carta Última – Parte 1”, deve ser discutida entre a Maria Isabel e a Maria Teresa, e entre a Maria Isabel e a Maria Velho, respectivamente.

Não há muitas explicações possíveis para o facto de a Maria Teresa aparecer, subitamente, como a autora mais provável dos textos. Uma das possíveis razões tem a ver com o facto de os textos de autoria desconhecida estarem “contaminados” por sinais de pontuação; aliás, foram estes textos que chamaram a atenção para a pontuação utilizada pelas autoras. Talvez o uso intensivo de sinais de pontuação nestes textos, tenha como consequência, o enviesamento do valor das permilagens calculadas à custa das frequências absolutas. Outra explicação que talvez possa ser avançada, tem a ver com a hipótese, já colocada, de uma das autoras ter efectuado uma revisão dos textos. Assim sendo, e perante os dados anteriores, a Maria Teresa surge desta vez como a revisora mais provável.

No entanto, a explicação mais lógica tem por base a variabilidade, ainda apresentada por alguns dos sinais considerados. Antes da análise efectuada, eliminaram-se o travessão, o ponto de interrogação e as aspas por apresentarem uma variabilidade elevada. No entanto, três dos sinais restantes, apresentam ainda uma variabilidade significativa, para uma das autoras; são eles: os dois pontos, os parêntesis e as pelicas. Na tentativa de tornar mais claras as conclusões anteriores, também se eliminaram do estudo estes sinais; ou seja, consideraram-se apenas aqueles que apresentam uma variabilidade inferior a dois.

Na tabela seguinte, apresentam-se os resultados obtidos.

Textos	Tendência
Primeira Carta I	Teresa
Primeira Carta II	Isabel/Teresa
Primeira Carta III	-
Primeira Carta Última – Parte 1	Velho
Primeira Carta Última – Parte 2	-
Primeira Carta Última – Parte 3	Teresa/Velho
Segunda Carta Última	Teresa/Velho
Terceira Carta Última	Teresa
Paz	Teresa
Lamento	Inconclusivo
Monólogo	Teresa/Velho
Escriturário	Inconclusivo
Tarefas	Isabel/Velho

Tabela 74: Atribuição de autoria com base nos sinais de pontuação com desvio-padrão inferior a dois

Repare-se na discrepância dos resultados obtidos, quando se resolveu "ignorar" os sinais de pontuação que apresentavam, ainda, um desvio-padrão elevado. Note-se que, agora, apenas se consideraram os sinais de pontuação com desvio-padrão inferior a dois.

A Maria Teresa aparece, agora, como a possível autora de apenas três textos — "Primeira Carta I", "Terceira Carta Última" e "Paz". Embora para a maioria dos textos, os resultados obtidos ainda indiquem a Maria Teresa como uma das possíveis autoras, a evidência de que esta seja a autora não é tão grande.

A Maria Velho aparece como a possível autora do texto, "Primeira Carta Última – Parte 2"; a autoria dos textos "Primeira Carta Última – Parte 3", "Segunda Carta Última" e "Monólogo" deverá ser discutida (segundo este critério) entre a Maria Velho e a Maria Teresa. A autoria do texto "Primeira Carta II" deve ser discutida entre a Maria Isabel e a Maria Teresa, e o texto "Tarefas" aparece como sendo um texto da Maria Isabel ou da Maria Velho. Para os restantes textos o estudo mostrou-se inconclusivo, devido à falta de dados, ou por os valores obtidos não apontarem de maneira convicta numa dada direcção.

Apesar do tipo e da frequência dos sinais de pontuação nos ter parecido uma variável promissora para ajudar a discriminar as autoras, tal não se verificou. A grande variabilidade apresentada pela frequência de utilização dos sinais, faz com que esta variável não seja das melhores e que o seu uso não seja muito recomendado. Mesmo assim, se se pretender utilizar a variável em trabalhos análogos a este, os sinais a utilizar devem ser os que apresentam a menor variabilidade.

3.6. Estudo do modo como as orações se encontram ligadas numa frase — coordenação e/ou subordinação

Numa tentativa, quase desesperada, de trazer alguma luz a este estudo, tendo em vista "arranjar" autoras para os textos desconhecidos, estudou-se ainda a forma como as orações se encontravam ligadas numa frase. As frases, de um modo geral, são constituídas por várias orações, orações essas que têm de estar ligadas por forma a formar a frase. Na língua portuguesa a ligação entre orações pode ser feita de dois modos, utilizando-se para tal, dois tipos de conjunções: as conjunções coordenativas, como o "e", "mas", "pois", etc. e as conjunções subordinativas, como o "se", "que",

"como", etc. Note-se, que uma frase pode ser constituída por mais do que um tipo de conjunções; em frases longas, por exemplo, algumas orações podem estar ligadas por coordenação e outras por subordinação.

Assim, para cada autora foi verificar-se a forma que estas mais utilizavam para ligar as orações dentro de uma frase; i.e., se utilizavam mais conjunções coordenativas ou mais conjunções subordinativas. Para tal, e para cada autora, seleccionaram-se aleatoriamente cem frases dos textos de autoria conhecida; o porquê da escolha das cem frases foi apresentado na subsecção 3.3, assim como o modo como as cem frases foram seleccionadas. Foi para as frases seleccionadas que se verificou o modo como as orações estavam ligadas.

Saliente-se o facto de o estudo do modo de ligação das orações, ter sido efectuado pela autora deste trabalho, sem outro recurso para além da "Nova Gramática do Português Contemporâneo" do Professor Lindley Cintra, [3]. O estudo da forma de ligação de orações não é trivial; existem frases para as quais os especialistas não conseguem chegar a um consenso, quanto ao modo como as orações se encontram ligadas. É por isso perfeitamente natural que haja quem não concorde com a classificação feita.

Fixados os parâmetros que serviram de base à classificação das frases, os resultados obtidos para os textos de autoria conhecida, apresentam-se na tabela seguinte.

Autora	**Coordenação**	**Subordinação**	**Coordenação e Subordinação**
Isabel	47%	34%	19%
Teresa	51%	29%	20%
Velho	37%	34%	29%

Tabela 75: Percentagem de "tipo de frase" encontrada nos textos conhecidos

Verifica-se que as três autoras utilizam mais os vocábulos coordenativos para ligar as orações, embora a Maria Velho utilize quase indistintamente os dois tipos de

vocábulos. Das três autoras, verifica-se ser a Maria Teresa que utiliza com maior ênfase os vocábulos coordenativos. De salientar ainda, o facto de a Maria Teresa e a Maria Isabel, utilizarem cerca de 20% de frases, onde as orações estão ligadas por coordenação e por subordinação; a Maria Velho utiliza cerca de 30% de frases deste tipo. A percentagem de frases onde as orações se encontram ligadas por coordenação e subordinação não é estranha, se se pensar que, por vezes, as autoras utilizam frases bastante longas.

De seguida, e para as frases dos textos de autoria desconhecida, efectuou-se um estudo análogo ao anterior. Isto é, foi verificar-se o modo como as orações se encontram ligadas nas frases.

Para se poder comparar os resultados obtidos nos textos de autoria desconhecida com os resultados obtidos nos textos de autoria conhecida, calculou-se, neste caso, a percentagem dos dados obtidos nos textos. E comparou-se a percentagem obtida nos textos desconhecidos, com a percentagem obtida nos textos conhecidos. Esta comparação foi efectuada, como até aqui, com base na distância mínima entre valores. As comparações efectuadas levaram às seguintes conclusões:

	Coordenação	**Subordinação**	**Coordenação e Subordinação**	**Tendência**
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
1ª Carta I	33.3%	50%	16.67%	
Tendência	**Velho**	**Isabel/Velho**	**Isabel**	**Isabel/Velho**

Tabela 76 :Resultados para a Primeira Carta I

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
1ª Carta II	25%	50%	25%	
Tendência	**Velho**	**Isabel/Velho**	**Velho**	**Velho**

Tabela 77: Resultados para a Primeira Carta II

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
1ª Carta II	30%	30%	40%	
Tendência	**Velho**	**Teresa**	**Velho**	**Velho**

Tabela 78: Resultados para a Primeira Carta III

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
1ª Carta Última-1	42.1%	15.8%	42.1%	
Tendência	**Isabel**	**Teresa**	**Velho**	**Inconclusivo**

Tabela 79: Resultados para a Primeira Carta Última – Parte 1

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
1ª Carta Última-2	23.8%	42.7%	33.3%	
Tendência	**Velho**	**Velho/Isabel**	**Velho**	**Velho**

Tabela 80: Resultados para a Primeira Carta Última – Parte 2

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
1ª Carta Última-3	27.28%	24.24%	48.48%	
Tendência	**Velho**	**Teresa**	**Velho**	**Velho**

Tabela 81: Resultados para a Primeira Carta Última – Parte 3

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
2ª Carta Última	52.44%	23.17%	24.39%	
Tendência	**Teresa**	**Teresa**	**Teresa**	**Teresa**

Tabela 82: Resultados para a Segunda Carta Última

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
3ª Carta Última	45.46%	27.27%	27.27%	
Tendência	**Isabel**	**Teresa**	**Velho**	**Inconclusivo**

Tabela 83: Resultados para a Terceira Carta Última

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
Paz	20%	65%	15%	
Tendência	**Velho**	**Velho/Isabel**	**Isabel**	**Isabel/Vel.**

Tabela 84: Resultados para a Paz

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
Monólogo	21.05%	31.58%	47.37%	
Tendência	**Velho**	**Teresa**	**Velho**	**Velho**

Tabela 85: Resultados para a Monólogo

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
Lamento	31.37%	45.1%	23.53%	
Tendência	**Velho**	**Velho/Isabel**	**Teresa**	**Velho**

Tabela 86: Resultados para a Lamento

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
Escriturário	29.41%	47.06%	23.53%	
Tendência	**Velho**	**Velho/Isabel**	**Teresa**	**Velho**

Tabela 87: Resultados para a Escriturário

Textos	Coordenação	Subordinação	Coordenação e Subordinação	Tendência
Isabel	47%	34%	19%	
Teresa	51%	29%	20%	
Velho	37%	34%	29%	
Tarefas	8.6%	22.8%	68.6%	
Tendência	**Velho**	**Teresa**	**Velho**	**Velho**

Tabela 88: Resultados para a Tarefas

Da análise das tabelas anteriores, conclui-se que a autora que aparece como sendo a mais provável é a Maria Velho, já que oito dos treze textos são agora atribuídos a esta autora. A Maria Teresa aparece agora, como sendo autora, unicamente, do texto "Segunda Carta Última"; e os textos "Primeira Carta I" e "Paz" vêm a sua autoria discutida entre a Maria Isabel e a Maria Velho. O estudo revelou-se inconclusivo para os textos "Terceira Carta Última" e "Paz".

Assim, e mais uma vez, o estudo da variável em causa não foi de encontro aos estudos já efectuados para as restantes variáveis, nem veio reforçar, de um modo geral, nenhuma das ideias já apresentadas. A única ideia que parece ganhar, cada vez mais consistência, é de que uma das autoras fez a revisão dos textos; nesta subsecção, a Maria Velho apresenta-se como a revisora mais provável.

Devido ao caracter, um tanto ou quanto subjectivo, que está na base da classificação do modo de ligação das orações, a variável estudada não pareceu ter um forte poder discriminativo. No entanto, não deve ser liminarmente eliminada pelas pessoas que pretendam efectuar um estudo desta natureza, pois o modo de ligação das orações nas frases é uma característica distintiva entre autores.

O estudo desta variável é um caso típico de um estudo que deve ser efectuado por estatísticos em parceria com linguistas, pois os casos duvidosos com que nos deparamos, provavelmente, seriam mais facilmente ultrapassados se se tivesse tido a colaboração de um especialista.

4. Conclusão

Antes de se tentar estabelecer uma autoria final, para os textos de autoria desconhecida, objectivo último deste trabalho, começa-se por chamar a atenção do leitor para algumas dificuldades sentidas no que respeita à definição das variáveis.

Depois de estudadas as dez variáveis apresentadas anteriormente, algumas conclusões podem ser retiradas deste trabalho. Como se disse no início, a definição das variáveis a utilizar em estudos desta natureza não é trivial, e, ao longo do mesmo, verificou-se que nem todas as variáveis utilizadas eram boas discriminadoras. Algumas das variáveis apresentadas, por exemplo, o tipo e frequência da pontuação utilizada pelas autoras, pareciam ser boas para as discriminar; porém, com o desenrolar do estudo, veio a verificar-se que algumas dessas variáveis não eram tão boas como à partida aparentavam ser. No caso da pontuação, a grande variabilidade associada à frequência de utilização dos sinais de pontuação, explica, em parte, o mau comportamento da variável, sendo por isso uma variável pouco recomendável. No entanto, as variáveis associadas à frequência das palavras não contextuais, ao comprimento de frase e ao comprimento de parágrafo revelaram-se as três variáveis com melhor comportamento, permitindo, de algum modo, discriminar as autoras. O estudo destas variáveis em trabalhos futuros, deve, por isso, ser aprofundado.

O estudo da frequência da algumas palavras “especiais” como o “pois”, e o “certo”, revelou-se, nalguns casos, importante, não sendo por isso de descartar o estudo de variáveis deste tipo. Foram estudados três vocábulos considerados “especiais”, tendo o seu estudo sido motivado pela constatação de que podiam ser utilizados em posições diferentes numa frase e/ou com sentidos diferentes. Como estes, outros vocábulos existem na língua portuguesa que podem e devem ser estudados, desde que ocorram nos textos em estudo.

Ao leitor deste trabalho, provavelmente, ocorreram outras variáveis que não as que são objecto deste estudo; por exemplo, o uso de sinónimos (embora as expressões “de um/de uma” possam ser consideradas sinónimos de “dum/duma”,

respectivamente), os tempos verbais utilizados, o uso de estrangeirismos, etc. No entanto, convém lembrar que o estudo de variáveis deste tipo requer um conhecimento profundo da língua portuguesa. Como tal, o estudo de variáveis deste tipo deve ser um trabalho de parceria entre estatísticos e estudiosos da língua portuguesa de modo a evitar as dificuldades (e provavelmente os erros) que a autora deste trabalho sentiu (e provavelmente cometeu) quando resolveu estudar, por exemplo, o tipo de ligação entre orações numa frase.

Refira-se que neste trabalho ainda se tentou estudar o uso de vocativos, pois em alguns dos textos desconhecidos o seu uso é frequente; no entanto, nos blocos de texto conhecido o uso de vocativos não foi detectado. Também se tentou estudar o recurso das autoras a estrangeirismos, pois em alguns dos textos conhecidos existiam não só estrangeirismos, mas também excertos de texto em inglês e francês. Mas, mais uma vez, nos textos desconhecidos seleccionados, as autoras não utilizaram estrangeirismos nem escreveram em inglês e/ou francês. Como tal, estas duas variáveis foram colocadas de parte, e como já foi dito, acabou por eliminar-se dos textos conhecidos os excertos escritos em inglês e/ou francês.

Como se pode constatar, uma das dificuldades deste tipo de trabalhos reside na "falta de dados"; i.e., aquando de estudo dos textos conhecidos (desconhecidos) depara-se com determinadas características que podem ser utilizadas para distinguir os autores, mas depois, quando se passa à análise dos textos desconhecidos (conhecidos), tais características estão ausentes. Por isso, é que em estudos desta natureza tudo o que seja contextual deve ser "posto de parte", pois não serve para distinguir os autores.

Outra das dificuldades encontradas neste trabalho, foi a escolha das técnicas estatísticas a utilizar, pois todas pareciam "demasiado elaboradas" para o que se pretendia efectuar. Depois de se saber quais as variáveis importantes para distinguir os autores, podem-se aplicar uma série de técnicas estatísticas, como a análise discriminante, ou utilizar a abordagem Bayesiana, de um modo similar à utilizada por Mosteller e Wallace nos papéis federais. O problema coloca-se quando ainda não foram seleccionadas as variáveis a utilizar nos estudos mais "elaborados". Como é que se chega à conclusão que, por exemplo, o comprimento de frase se deve utilizar numa análise discriminante, se não se fizer a mínima ideia se esta variável

funciona bem ou mal, i.e., se permite distinguir os autores. Foi no estudo inicial das variáveis que se sentiu "falta" de técnicas estatísticas, que fossem além da comparação do valor observado com o valor da média ou da mediana.

Provavelmente, as técnicas que foram utilizadas neste trabalho são as que devem ser utilizadas na fase inicial de qualquer trabalho deste tipo. E todo o processo de atribuição de autoria, deve começar por uma fase inicial análoga à aqui efectuada, para desta forma se chegar às variáveis que devem ser estudadas com maior pormenor, e com recurso a técnicas estatísticas mais "elaboradas" possuindo maior grau de precisão.

Antes de se passar a uma atribuição de autoria para os textos desconhecidos, convém mais uma vez referir a necessidade sentida, ao longo deste trabalho, de ter disponível a ajuda de uma pessoa cuja área de estudo fosse a língua portuguesa. Se tal tivesse acontecido, algumas das dúvidas e dificuldades sentidas, teriam tido outra resposta. Além do mais, existe, na área da linguística, software que pode auxiliar o trabalho, nomeadamente no que diz respeito às contagens, e a colaboração com pessoas da área, provavelmente, vem acrescida com o acesso mais fácil a esse software.

De seguida, vão-se apresentar as conclusões finais, a que se chegou, sobre a autoria das textos desconhecidos. Começa-se por apresentar, na tabela seguinte, um resumo das atribuições de autoria feitas aquando do estudo de cada uma das variáveis.

Textos	Palavras	Frases			Parágrafos		Pontu.	Certo	Depois	Dum	Pois	Não
		Comp.	Assi.	Tipo	Comp.	Assi.						
Primeira Carta I	-	Velho	Velho	Isa/Velho	-	-	Teresa	Isa/Velho	-	-	-	-
Primeira Carta II	Ter/Velho	Teresa	Teresa	Velho	Isabel	Teresa	Isa/Teresa	-	-	Ter/Velho	-	-
Primeira Carta III	Velho	Isabel	Velho	Velho	-	-	-	-	Velho	-	-	Velho
Primeira Última – Parte 1	Isabel	Velho	Velho	-	Velho	-	Velho	-	-	-	-	-
Primeira Última - Parte2	Isabel	Velho	Velho	Velho	Isabel	Teresa	-	Isa/Velho	-	-	-	Velho
Primeira Última – Parte 3	Isabel	Velho	Velho	Velho	Velho	Isabel	Ter/Velho	Isa/Velho	Velho	-	Velho	-
Segunda Última	Isabel	Teresa	Isabel	Teresa	Teresa	Teresa	Ter/Velho	-	Isabel	Ter/Velho	-	-
Terceira Última	Isabel	Velho	Velho	-	Isabel	Teresa	Teresa	-	Velho	-	-	Isabel
Paz	Ter/Velho	Velho	Velho	Isa/Velho	Isabel	Teresa	Teresa	-	-	-	-	Teresa
Lamento	Velho	Velho	Velho	Velho	Isabel	Isabel	-	-	Teresa	-	Velho	Isabel
Monólogo	Isabel	Isabel	Teresa	Velho	-	-	Ter/Velho	-	Velho	Ter/Velho	-	-
Escriturário	Velho	Isabel	-	Velho	Isabel	Teresa	-	-	-	Ter/Velho	-	Isabel
Tarefas	Isabel	Isabel	Velho	Velho	Isabel	Velho	Isa/Velho	-	Teresa	Ter/Velho	Velho	Velho

Tabela 87 : Resumo das atribuições de autoria que o estudo de cada variável conduziu

De seguida, analisam-se os resultados apresentados na tabela anterior, e apresenta-se uma interpretação dos mesmos.

Para a “Primeira Carta I” apenas se encontraram dados para três das variáveis estudadas — comprimento de frase, tipo de frase e pontuação. O comprimento de frase “apontou” a Maria Velho como a possível autora do texto, o tipo de frase indicou a Maria Isabel ou a Maria Velho como as possíveis autoras, e a pontuação utilizada a Maria Teresa. Como a Maria Velho é das três autoras a que apresentou mais referências, aposta-se nesta autora, como a mais provável autora do texto “Primeira Carta I”.

A maioria das variáveis estudadas para a “Primeira Carta II”, indicou a Maria Teresa como a possível autora do texto. O comprimento de frase (tanto considerando o valor médio do comprimento de frase, como o valor do coeficiente de assimetria), e o comprimento de parágrafo (quando se considerou o coeficiente de assimetria) indicam a Maria Teresa como a autora mais provável do texto. Esta autora é também, uma das autoras indicadas pelas variáveis: frequência das palavras não contextuais, pontuação e frequência de utilização das expressões “de um/de uma” e “dum/duma”.

As variáveis analisadas para a “Primeira Carta III”, indicam a Maria Velho como a possível autora deste texto. Facto que está em desacordo com a hipótese colocada no início deste trabalho, de que as três autoras se apresentaram e se despediram “desta aventura”; pois a Maria Velho aparece, também, como a mais provável autora do texto “Primeira Carta I”. Como as evidências de que a Maria Velho é a autora da “Primeira Carta III”, são mais fortes do que as evidências de que é ela a autora da “Primeira Carta I”, aposta-se mais na Maria Velho como autora da “Primeira Carta III”, e mantendo a hipótese inicial, i.e., as três autoras apresentaram-se e despediram-se, a Maria Isabel é possivelmente a autora da “Primeira Carta I”.

A hipótese, também colocada no início deste trabalho, de que a autora das três partes que constituem a “Primeira Carta Última” é a mesma, sai reforçada pelos resultados obtidos. Os resultados obtidos apontam a Maria Velho como a provável autora dos três textos referidos.

A possível autora da “Segunda Carta Última”, aparece como sendo a Maria Teresa. Note-se no entanto que, quer a Maria Isabel quer a Maria Velho, também são referidas como as possíveis autoras do texto, mas as evidências mais fortes vão no sentido da Maria Teresa.

Para “Terceira Carta Última” regista-se um “empate” entre a Maria Isabel e a Maria Velho, como as possíveis autoras do texto. No entanto, como a Maria Velho aparece como a autora mais provável das três partes que constituem a “Primeira Carta Última”, aposta-se mais na Maria Isabel como a possível autora da “Terceira Carta Última”, partindo do princípio de que a hipótese de que as três autoras se apresentaram e se despediram, é verdadeira.

Para os restantes cinco textos, a Maria Isabel aparece como a possível autora do texto “Escriturário”, a Maria Teresa talvez seja a autora do texto “Paz”, e a Maria Velho a possível autora dos textos “Lamento”, “Monólogo” e “Tarefas”.

Apresenta-se, na tabela seguinte, um resumo do que atrás se disse.

Texto	**Possível autora**
Primeira Carta I	Maria Isabel
Primeira Carta II	Maria Teresa
Primeira Carta III	Maria Velho
Primeira Carta Última – Parte 1	Maria Velho
Primeira Carta Última – Parte 2	Maria Velho
Primeira Carta Última – Parte 3	Maria Velho
Segunda Carta Última	Maria Teresa
Terceira Carta Última	Maria Isabel
Paz	Maria Teresa
Lamento	Maria Velho
Monólogo	Maria Velho
Escriturário	Maria Isabel
Tarefas	Maria Velho

Tabela 88: Atribuição de autoria feita como base no estudo efectuado

Os resultados obtidos sobre a possível autoria dos textos desconhecidos, de um modo geral, não são estranhos. Apenas para os textos “Primeira Carta III” e “Terceira Carta Última”, os resultados obtidos não vão ao encontro das hipóteses

formuladas no início deste trabalho. Facto que não parece ser suficiente para colocar em dúvida as hipóteses formuladas, embora também não prove a veracidade das mesmas. Note-se ainda, que a Maria Teresa aparece, apenas, como a possível autora dos textos “Primeira Carta II”, “Segunda Carta Última” e “Paz”, o que está de acordo com hipótese formulada de que esta autora escreveu os poemas que constam nas “Novas Cartas Portuguesas” e, como tal, escreveu menos textos em prosa, deixando estes para as outras duas autoras.

De salientar, também, o facto de nenhum dos textos desconhecidos ter apenas uma autora como a mais provável, já que, para quase todos os textos, as três autoras são referidas como a possível autora.

Este facto vem reforçar a ideia, já apresentada, de que provavelmente uma das três autoras fez um trabalho de revisão dos textos. Se tal se verificou, a Maria Isabel e a Maria Velho surgem como as revisoras mais prováveis, pois são as que mais vezes são referidas como as possíveis autoras dos textos. Para todos os textos desconhecidos surge sempre uma ou mais variáveis, que indicam que a possível autora dos textos é a Maria Isabel e/ou a Maria Velho. No entanto, se a revisora dos textos tivesse sido a Maria Teresa, tal facto não causaria grande surpresa, pois aquando do estudo da pontuação, esta autora apareceu como a possível autora da maioria dos textos. Além disso, se se pensar que quando alguém faz a revisão de um texto, o que mais é alterado é a pontuação, a hipótese de ter sido a Maria Teresa a efectuar a revisão dos textos ganha outra dimensão.

As conclusões apresentadas anteriormente não serão, provavelmente, as únicas que podem ser retiradas do estudo efectuado. É natural que devido à grande subjectividade que está subjacente a este trabalho, que outras pessoas, em face dos mesmos resultados, tivessem tirado outras conclusões. As conclusões apresentadas pareceram ser as menos polémicas e as mais lógicas, mas estamos abertos a críticas, e comentários. Sugestões serão sempre bem-vindas.

Um trabalho futuro, como já foi dito, passa por um estudo mais aprofundado de algumas das variáveis anteriores — frequência das palavras não contextuais, comprimento de frase e comprimento de parágrafo — e pela pesquisa de outras

variáveis importantes na definição e identificação do estilo dos autores, de modo a confirmar ou não as conclusões anteriores.

Comentário Final

Uma semana depois de defender a dissertação de mestrado recebi uma carta da Drª Maria Isabel Barreno na qual esta comentava os resultados obtidos com este trabalho. Assim, e no total, conseguimos acertar com a autoria de cerca de 50% dos textos em estudo.

Referências

[1] M. I. Barreno, “Os Outros Legítimos Superiores”, 2ª edição (1993), Editorial Caminho.

[2] M. I. Barreno, M. T. Horta, M. V. da Costa, “Novas Cartas Portuguesas”, 2ª Edição (1974), Publicações Futura.

[3] C. Cunha, L. Cintra, “Nova Gramática do Português Contemporâneo”, 13ª edição (1997), Edições João Sá da Costa, L.da.

[4] M. V. da Costa, “Maina Mendes”, 3ªedição, (1993), Publicações Dom Quixote.

[5] D. C. Hoaglin, F. Mosteller, J. W. Tukey, “Exploring Data Tables, Trends, and Shapes”, (1985), John Wiley & Sons.

[6] D. C. Hoaglin, F. Mosteller, J. W. Tukey, “Análise Exploratória de Dados Técnicas Robustas”, (1992), Edições Salamandra.

[7] M. T. Horta, “Ambas as Mãos Sobre o Corpo”, 3ª edição (1970), Publicações Europa-América.

[8] F. Mosteller, David L. Wallace, “Applied Bayesian and Classical Inference — The Case Of The Federalist Papers”, Second Edition (1983), Springer-Verlag.

[9] F. Mosteller, J. W. Tukey, “Data Analysis And Regression a Second Course in Statistics”, (1997), John Wiley & Sons.

[10] B. J. F. Murteira, “Estatística Descritiva”, (1993), McGraw Hill.