DOI: http://dx.doi.org/10.5007/1807-9288.2013v9n1p24

ATRIBUIÇÃO DE AUTORIA UTILIZANDO ANÁLISES ESTATÍSTICAS: UMA EXPERIÊNCIA COM A RELAÇÃO ABREVIADA[1]

Emanoel César Pires de Assis[*]

RESUMO: O presente artigo pretende, a partir de análises estatísticas e estilométricas, utilizando o Hyperbase, software de tratamento de textos que, através do reconhecimento das particularidades estruturais da língua, mapeia as categorias textuais e as transforma em dados estatísticos, realizar uma aproximação entre a *Relação Abreviada*, texto de autoria desconhecida, e seus possíveis autores: o Marquês de Pombal e Basílio da Gama. Para tanto, procedimentos e análises envolvendo a pontuação dos textos, as altas frequências, as palavras funcionais, a repetição de passagens e outros foram realizados e nos deram fortes subsídios para acreditar que a *Relação Abreviada* pode ter sido escrita pelo Marquês de Pombal.

Palavras-Chave: Atribuição de autoria. Estatística. Hyperbase. Marquês de Pombal. Basílio da Gama.

> *É a linguagem que delimita o campo das significações, das interpretações possíveis de uma obra literária. É ela que dá identidade, possibilitando que todos possam reconhecer uma obra como a mesma, e não outra. É no trabalho de linguagem, no modo como o autor lida com esse instrumento, que está sua arte*
> (FREITAS, 2007).

Através de procedimentos estatísticos, tentaremos, ao longo do nosso texto, colocar à prova a hipótese de que a *Relação Abreviada*[2] foi escrita por Sebastião José de Carvalho e Melo, o Marquês de Pombal. Para isso, utilizaremos o Hyperbase[3], software de tratamento de textos que, através do reconhecimento das particularidades estruturais da língua, mapeia as categorias textuais e as transforma em dados estatísticos.

[1] Agradeço ao professor Carlos Maciel, pela sua paciência e disponibilidade, e à Deise Freitas pelas dúvidas tiradas.
[*] Universidade Federal de Santa Catarina. Imeio: lordemanoel@hotmail.com.
2 O nome completo da obra é: Relação abreviada da república que os religiosos jesuítas das províncias de Portugal e Espanha estabeleceram nos domínios ultramarinos das duas monarquias, e da guerra que neles têm movido e sustentado contra os exércitos espanhóis e portugueses; formado pelos registros das secretarias dos dois respectivos principais comissários e plenipotenciário, e por outros documentos autênticos. Por questões de economia, quando nos referirmos à obra, utilizaremos apenas o título Relação Abreviada, ou apenas Relação.
[3] O Hyperbase foi desenvolvido por Etienne Brunet, da Universidade de Nice.

Em uma rápida pesquisa pela internete, pudemos verificar que algumas dezenas de sítios dão a Basílio da Gama a autoria da *Relação Abreviada*. Porém, estudos mais sérios como o de Vânia Pinheiro Chaves (2000), J. Lucio de Azevedo (1922) e alguns artigos de fontes mais confiáveis afirmam que a obra é anônima.

Sem revelar fontes e critérios, Sacramento Blake (1883) também coloca Basílio da Gama como o suposto autor da *Relação Abreviada*. Acreditamos que muitos dos textos que atribuem a autoria da *Relação* a Basílio estejam fundados na semelhança temática existente entre a obra e *O Uraguai*. Os dois textos possuem uma clara ideologia antijesuítica e foram publicados com uma pequena diferença temporal entre si. O primeiro em 1757 e o segundo em 1769, ou seja, há um intervalo de 12 anos entre as publicações, o que para a época não caracteriza uma grande distância temporal.

Não encontramos nenhum texto com o mínimo de rigor científico que atribua a autoria da *Relação Abreviada* ao Marquês de Pombal. Porém, objetivando relativizar os dois possíveis autores, decidimos acrescentar à base de textos que seriam analisados, obras de Basílio da Gama e textos do Marquês de Pombal.

Encontrar obras atribuídas a Basílio da Gama não nos foi difícil. O autor deixou algumas dezenas de poemas escritos, porém textos de autoria do Marquês de Pombal já são mais raros e de pouco conhecimento acadêmico. Felizmente, tivemos acesso a dois volumes dedicados exclusivamente à correspondência e demais escritos do Conde de Oeiras, como Sebastião José costumava assinar antes de ser conhecido como o Marquês de Pombal. O acesso aos textos se deu através de uma cópia digitalizada do primeiro volume, publicado em Lisboa, em 1861. O documento em formato impresso encontra-se na biblioteca da Universidade de Toronto, no Canadá.

De posse de amostras textuais dos dois possíveis autores, partimos para o tratamento dos textos, revisão, adequação da ortografia e criação da nossa

base textual. Todos esses procedimentos serão pormenorizados agora na seção dedicada aos procedimentos de normatização das obras.

Preparação dos textos

Para a feitura desse trabalho, tivemos que passar por algumas etapas metodológicas. Assim, um critério minimamente rigoroso de tratamento dos textos analisados teve que ser seguido. Tal critério, ou critérios, objetivava assegurar, com um grau de confiabilidade aceitável, uma boa interpretação dos dados retirados a partir da inserção dos textos no Hyperbase.

Primeiramente, buscamos por uma fonte confiável do texto que seria posto à prova de autoria. Para tanto, retiramos o texto da *Relação* no sítio literaturabrasileira.ufsc.br, de responsabilidade do Núcleo de Pesquisa em Informática Linguística e Literatura- NUPILL, e que usou como fonte primária, o texto publicado na Revista do Instituto Histórico e Geográfico Brasileiro. Optamos pelo texto encontrado no sítio eletrônico, por ele já estar em um formato que poderia ser facilmente exportado para o Hyperbase e por já ter tido sua ortografia atualizada. Enfatizamos que mesmo assim o texto ainda passou por um processo de busca de imperfeições e disparidades em relação à edição da revista do Instituto.

É interessante notar que o sítio do NUPILL atribui a autoria da *Relação* tanto a Basílio quanto a Pombal. O que nos deixou mais inquietos ainda quanto à autoria e nos fez buscar subsídios que pudessem revelar traços autorais e tentar desfazer a relação dicotômica existente.

Depois de tomados esses cuidados iniciais e de posse do texto “limpo”, passamos ao trabalho de buscar uma base de textos de Basílio da Gama que pudesse servir como parâmetro de estilo do poeta. Assim, elencamos e juntamos em um único arquivo, uma gama de textos de autoria de Basílio. A saber: *A Declaração Trágica; Os Campos Elísios*; *Canto Único ao Marquês de Pombal*; *Epitalâmio às Núpcias da Senhora Dona Maria Amália, Filha do Marquês de Pombal*; *Glosa à Quadra do Duque de Lafões*; *Lenitivo da*

Saudade; *O Entrudo*; *Ode a Vasco da Gama*; *Ode ao Conde da Cunha*; *Ode ao Reio José*; *Quitubia* e *O Uraguai* (sem a notas).

Ao todo, nosso *corpus* de textos de Basílio contava com 20.343 ocorrências e 4.480 formas[4]. Optamos por escolher os textos de maior extensão, porque eles nos dariam uma maior representatividade do traço estilístico de Basílio da Gama. Dessa forma, deixamos de fora alguns textos mais conhecidos pelo público leitor, porém de menor tamanho lexical.

Vale mencionar que as obras de Basílio da Gama também foram retiradas do sítio do NUPILL e têm como fonte de referência *Obras Poéticas de Basílio da Gama*, editada pela Edusp. Os textos presentes na biblioteca virtual do NUPILL passaram pelos procedimentos de revisão técnica e, quando necessário, atualização ortográfica. O que nos dá uma certa homogeneidade no padrão com que os textos foram tratados e nos faz fugir de imperfeições resultantes de textos que foram retirados de fontes diferentes entre si. Mesmo assim, não nos custa mencionar mais uma vez, tanto os textos de Basílio quanto os de Pombal passaram pelos mesmos procedimentos de normatização e padronização realizados por nós.

No que diz respeito à padronização gráfica dos textos, tivemos que fazer modificações principalmente nos textos de Basílio da Gama, uma vez que estávamos comparando textos de gêneros diferentes, e a poesia de Basílio, visando a métrica dos versos, possuía características típicas do gênero poético. Como do outro lado tínhamos um texto em prosa, resolvemos uniformizar a grafia entre os textos. Como exemplo podemos citar os vários casos de contrações e elisões que o poeta fez durante a escrita de suas obras. Assim, em casos como “Co'a cândida Justiça a Paz dourada”, modificamos para *Com a cândida Justiça a Paz dourada*.

[4] Entende-se por forma qualquer unidade que esteja presente no texto, ou seja, o conceito engloba tanto as palavras quanto os sinais de pontuação. Ocorrências são todas as unidades do texto, as formas e suas repetições.

Essas modificações gráficas não interferem no sentido do texto, muito menos alteram a estrutura sintática das frases. Ao contrário, elas nos ajudam a fazer com que os textos obedeçam a um certo padrão. Outros casos também foram observados:

Texto Original	Texto Modificado
Foge d'ao redor dele a iníqua Guerra	Foge de ao redor dele a iníqua Guerra
Pinte o terror dos olhos toda a desordem d'alma	Pinte o terror dos olhos toda a desordem da alma
Projetos horrorosos, que forma um'alma ímpia	Projetos horrorosos, que forma uma alma ímpia
O público embebido c'oa trágica grandeza	O público embebido com a trágica grandeza
Voss'alma porventura toda jamais se entrega?	Vossa alma porventura toda jamais se entrega?
Parece que inda pedem aos Céus o Herói que el'ama	Parece que inda pedem aos Céus o Herói que ela ama
Co's não vingados ossos dos parentes	Com os não vingados ossos dos parentes

Tabela 1 – Modificações feitas nos textos poéticos de Basílio da Gama.

Esses são apenas alguns dos casos em que tivemos que fazer uma intervenção gráfica, além de casos de supressão de vogais, também padronizamos a grafia das obras no que diz respeito aos vocábulos do tipo: oiro; loiro; tesoiros; oiça, etc. Modificados para: ouro; louro; tesouros e ouça, respectivamente. As modificações foram feitas com extrema cautela e sempre observando se o texto não sofreria alguma mudança drástica, o que corromperia o texto original e colocaria por terra nosso trabalho.

Após a padronização dos textos, passamos à tarefa de exportá-los para o formato .txt, formato de arquivo em que somente a parte textual é mantida e, por isso mesmo, único formato que o Hyperbase aceita.

A primeira base estava pronta, ela era composta pelos poemas de Basílio da Gama e por um número significativo de cartas e documentos escritos pelo Marquês de Pombal, que totalizaram 29.786 ocorrências e 5.165 formas. Um número maior de textos de autoria do Marquês de Pombal poderia ser

acrescentado, já que contávamos com dois volumes de documentos de sua autoria. Porém, no processo de exportação dos textos em formato .pdf para o formato .doc e em seguida para .txt, muitos dados se corrompem devido ao fato de que o programa utilizado para a conversão não se demonstrar ser inteiramente eficaz. O sistema, muitas vezes, não reconhece perfeitamente os caracteres e os exporta de forma errada.

Com isso, trechos em que no original havia uma vírgula, o sistema reconhecia como um ponto, ou vice-versa, e onde havia um *e* o sistema reconhecia como um *c*. Assim, toda uma gama de palavras sofria alteração e um processo de revisão era necessário. Como o .pdf foi feito a partir de um exemplar que data de 1861, as páginas já haviam sofrido um desgaste por conta do tempo e qualquer pequena partícula era reconhecida como uma vírgula, um ponto ou um acento. Dessa forma, optou-se por ter uma quantidade de texto que satisfizesse nossas exigências, e que fosse cautelosamente analisada e confrontada com o seu original.

Quanto à seleção dos textos de Pombal, seguiu-se o critério da ordem em que os textos apareciam. Suprimindo, sempre, a assinatura do autor ao final dos textos, bem como as datas em que os documento foram escritos. Uma vez que tais dados eram dispensáveis à nossa análise.

Após as revisões e padronizações, contávamos com três textos: *A Relação Abreviada* com 7.729 ocorrências[5]; os poemas de Basílio da Gama com 20.363 ocorrências e os documentos escritos por Pombal com 29.786 ocorrências. Ao observar nossa base textual, percebemos que tínhamos um problema quanto aos gêneros que estavam sendo analisados. De um lado havia a prosa (*Relação Abreviada* e os escritos de Pombal) e de outro lado tínhamos a poesia (poemas de Basílio da Gama). As comparações entre os textos poderiam ser feitas, porém a distinção de gênero entre eles poderia afetar os

[5] Retiramos do texto da Relação, os documentos I, II, III e IV, uma vez que a autoria deles, supostamente, não é da mesma pessoa.

dados. Assim, acrescentamos à base as notas presentes no *Uraguai*[6], já que elas estão em prosa e, assim, teríamos como fazer comparações com textos do mesmo gênero.

Após todos esses passos, começamos a inserir os textos no programa e a buscar características de estilo que pudessem aproximar a *Relação Abreviada* a um de seus dois supostos autores. Os resultados foram satisfatórios, mas ainda sentíamos que a construção de um argumento que pudesse ser validado no que diz respeito à autoria do texto anônimo estaria inconsistente se fechássemos nossa análise apenas aos textos já citados.

Portanto, tivemos que criar mais bases e acrescentar mais textos para verificar se a obra anônima continuava a se comportar da mesma maneira. À base inicial, com apenas quatro textos – *A Relação Abreviada*, os poemas de Basílio, as notas presentes no *O Uraguai* e os documentos escritos por Pombal – acrescentamos: o segundo volume dos *Autos da Devassa*[7]; *Marília de Dirceu*, de Tomás Antônio Gonzaga e as *Cartas Chilenas,* obra de autoria materialmente ainda não comprovada.

Os textos inseridos no programa possuem as seguintes características:

Obras	Ocorrências	Formas	Abreviação
Poemas de Basílio da Gama	20.363	4.480	Basíl
Textos do Marquês de Pombal	29.786	5.165	Pom
Relação Abreviada	7.729	1.862	Rela
Notas de O Uraguai	4.301	1.479	Not
II Volume dos Autos da Devassa	124.215	5.784	Aut
Marília de Dirceu	28.203	4.452	Marí
Cartas Chilenas	32.553	5.620	CarChi

[6] Vale mencionar que as notas, em prosa, também são de autoria de Basílio da Gama.
[7] O extenuante trabalho de revisar e transformar o enorme arquivo em .txt foi trabalho do professor Carlos Maciel, especialista, entre outras coisas, em estatística literária.

Tabela 2 – Número de ocorrências e formas por obra.

A escolha das obras se deve à proximidade histórica existente entre elas. Considerando que o léxico sofre alterações ao longo dos anos, palavras são inseridas e outras deixam de ser utilizadas, decidimos criar uma base textual que pudesse servir de parâmetro de comparação entre a *Relação Abreviada* e as obras dos possíveis autores. Textos com datas de publicação muito distantes poderiam causar interpretações enganosas.

Efetuado esses procedimentos técnicos, partimos para a extração e análise dos dados obtidos. Fizemos inúmeros experimentos e eles serão demonstrados a partir de agora.

Análise dos dados

O Hyperbase nos permite realizar inúmeras aproximações entre os textos em análise, contudo, haja vista a extensão e o caráter introdutório do nosso trabalho, focaremos nossa análise em alguns aspectos mais específicos. O primeiro dado retirado da nossa base textual diz respeito à frequência predominante dos diversos tipos de pontuação. A partir da análise, os textos se comportaram da seguinte maneira:

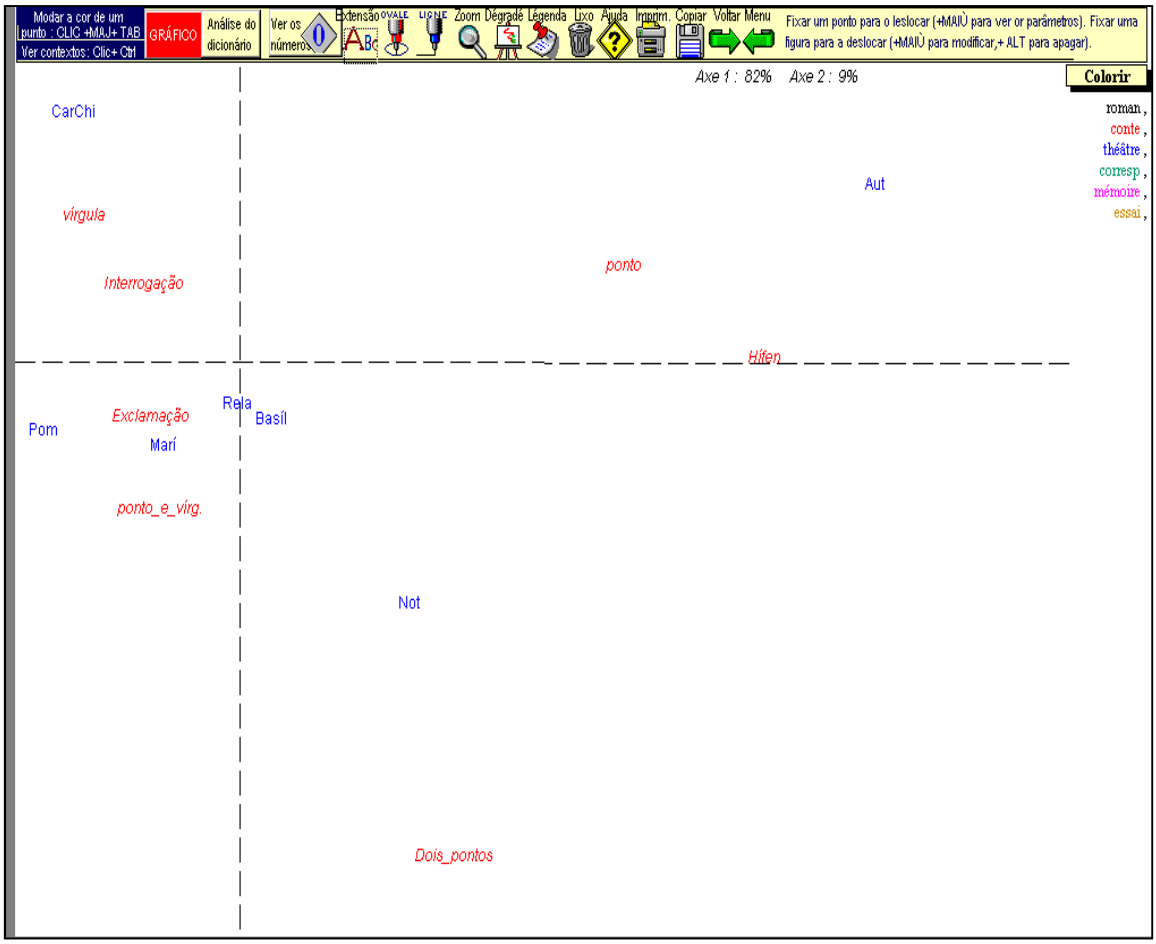

Gráfico 1 – Análise fatorial dos sinais de pontuação.

Observando a análise fatorial acima, podemos perceber a aproximação entre as obras e os sinais de pontuação. É possível verificar que a *Relação Abreviada* situa-se quase que no centro dos eixos, distanciando-se mais significativamente dos *Autos da Devassa*.

No tocante à pontuação, a *Relação* aproxima-se mais dos poemas de Basílio da Gama, mesmo assim, podemos perceber que a obra está situada entre a linha divisória de dois quadrantes, o que nos faz entender que a pontuação forte do texto é marcada por vírgulas, sinais de interrogação e exclamação e pelo ponto e vírgula, sinais de forte presença também nos textos de Pombal e em *Marília de Dirceu*.

No que diz respeito à distribuição dos sinais de pontuação ao longo dos textos, verificamos as seguintes quantidades:

Basíl	Pom	Rela	Not	Aut	Marí	CarChi	Sinal de Pontuação
1708	2779	560	274	9720	2805	3908	,
50	5	0	0	19	240	153	!
737	467	103	219	5072	931	896	.
148	158	41	106	776	211	126	:
163	240	38	17	586	511	266	;
31	6	4	1	76	149	178	?
165	227	79	75	1639	266	219	-

Tabela 3 – Distribuição dos sinais de pontuação ao longo das obras.

Como os textos têm extensões diferentes, calcular a aproximação entre eles a partir da quantidade de vezes ou da média em que determinada forma aparece geraria dados equivocados, na medida em que a proporção das palavras, sinais ou demais elementos textuais está intrinsecamente relacionada com a extensão do texto. Visando obter resultados que nos dessem uma maior confiabilidade, resolvemos utilizar a análise em árvore, calculada não a partir da média, mas a partir do desvio em relação ao *corpus*.

Sobre a análise em árvore, Xuan Luong (apud FREITAS, 2007, p. 91-92) afirma:

> L'analyse factorial des correspondances (AFC) offre, certes, répresentation simultané et directions principales, mais ne donne pas le détail et les hiérarchies entre proximités. Elle peut même conduire à des erreurs d'interprétations (projections, voisines sur un plan, de points eloignés dans l'espace), ce qui est impossible dans une arbre (où les distances se lisent en suivant les arcs et non pas en mesurant visuellement des distances entre feuilles.

Além de mais completa, a análise em árvore possui uma leitura mais fácil, e as interpretações a partir dos gráficos se dão de maneira mais eficiente. A mesma análise feita anteriormente, só que agora em árvore, gerou-nos o seguinte gráfico:

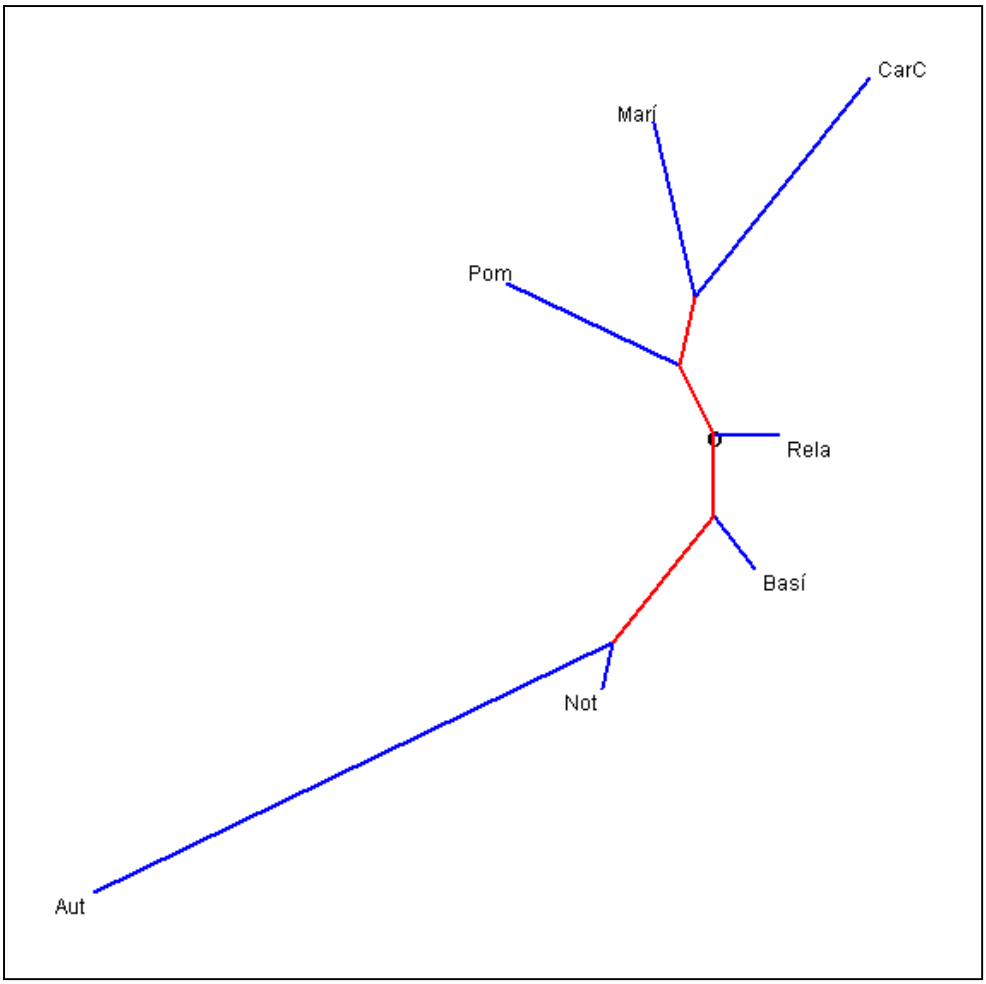

Gráfico 2 – Análise em árvore dos sinais de pontuação.

Interessantemente, *A Relação Abreviada* encontra-se entre os textos de Pombal e os poemas de Basílio da Gama. O que já nos antecipa que há uma relação existente entre a obra analisada e os dois possíveis autores.

A análise da pontuação nos permite visualizar importantes elementos estilísticos dos textos. A saber: a extensão das frases e o ritmo do texto. Para o criador do Hyperbase:

> La ponctuation pourrait constituer um domaine privilegié de la linguistique quantitative. Elle permet en effet d'échapper au cercle étroit du mot et d'ouvrir une perspective sur la phrase et le rythme du discours. Tant que la reconaissance du sens – qui conditionne dans une large part celle de la syntaxe – n'aura pas trouvé une solution acceptable dans le traitement automatique des textes, la ponctuation restera l'um des seuls accès au supralexical dont on puisse tirer parti (BRUNET, 1981, p.1).

Como menciona Brunet, o tratamento automático de textos isola as palavras de seus contextos, impossibilitando uma relação semântica mais abrangente. É por isso que a análise da pontuação é um importante auxílio no estudo de estilos autorais. Para a pesquisadora Deise Freitas (2007, p. 119): "Outra vantagem da análise da pontuação é que se trata de uma categoria com poucas variações: não são muitos os sinais e, além disso, eles não sofrem nenhum tipo de flexão". Contudo, como estamos trabalhando com obras de gêneros diferentes, e sendo a pontuação um elemento fortemente ligado ao gênero, decidimos efetuar mais análises.

Após o primeiro experimento com as pontuações, o que nos deixou curiosos por mais informações, buscamos relacionar os textos a partir das altas frequências, ou seja, as 100 formas com maior aparecimento ao longo de todo o *corpus*.

As altas frequências nos permitem ter uma ideia da temática tratada no texto (substantivos, verbos, adjetivos, etc.), bem como de sua estrutura (preposições, pronomes, conjunções, interjeições, etc.). Vejamos quais as formas mais encontradas no *corpus*:

Ordem	**Freq.**	**Forma**	18	1575	com	35	792	mesmo
1	21829	,	19	1574	:	36	791	sua
2	10136	.	20	1452	por	37	696	foi
3	9330	que	21	1422	para	38	682	à
4	7545	e	22	1330	mais	39	651	/
5	6637	de	23	1308	ele	40	629	como
6	6575	o	24	1302	ao	41	610	ou
7	5633	a	25	1115	um	42	596	é
8	4581	-	26	1103	José	43	595	tinha
9	3917	se	27	1098	dito	44	559	já
10	3023	do	28	1004	na	45	553	das
11	2267	os	29	982	no	46	550	me
12	2073	em	30	941	dos	47	529	uma
13	2024	não	31	923	(	48	487	este
14	1981	da	32	914	)	49	475	eu
15	1821	;	33	802	seu	50	467	!
16	1691	as	34	794	respondente	51	464	respondeu
17	1639	lhe						

52	457	também	69	374	sem	86	290	cel
53	445	?	70	354	tem	87	287	todos
54	444	aos	71	346	vila	88	286	sobre
55	442	nem	72	338	nas	89	283	Francisco
56	434	quem	73	336	senhor	90	280	logo
57	430	só	74	335	são	91	279	nesta
58	430	1	75	335	desta	92	276	dias
59	416	quando	76	334	tudo	93	272	perguntas
60	413	nos	77	315	Antônio	94	271	aquele
61	405	casa	78	309	assim	95	270	bem
62	401	mas	79	308	onde	96	266	ainda
63	397	pelo	80	301	sendo	97	265	tempo
64	392	era	81	297	oliveira	98	262	ser
65	386	seus	82	297	alguma	99	260	verdade
66	376	perguntado	83	294	2	100	259	Depois
67	376	ministro	84	291	esta			
68	376	“	85	290	escrivão			

Tabela 4 – Altas frequências.

Como o sistema busca as 100 formas mais frequentes, era esperado que nos primeiros lugares estivessem os sinais de pontuação, bem como as palavras funcionais mais comuns. Os verbos, substantivos, adjetivos, etc. – palavras que se opõem às funcionais, já que trazem consigo uma força semântica maior – aparecem de forma mais escassa e em posições mais inferiores.

Analisados os dados das altas frequências, temos o seguinte gráfico:

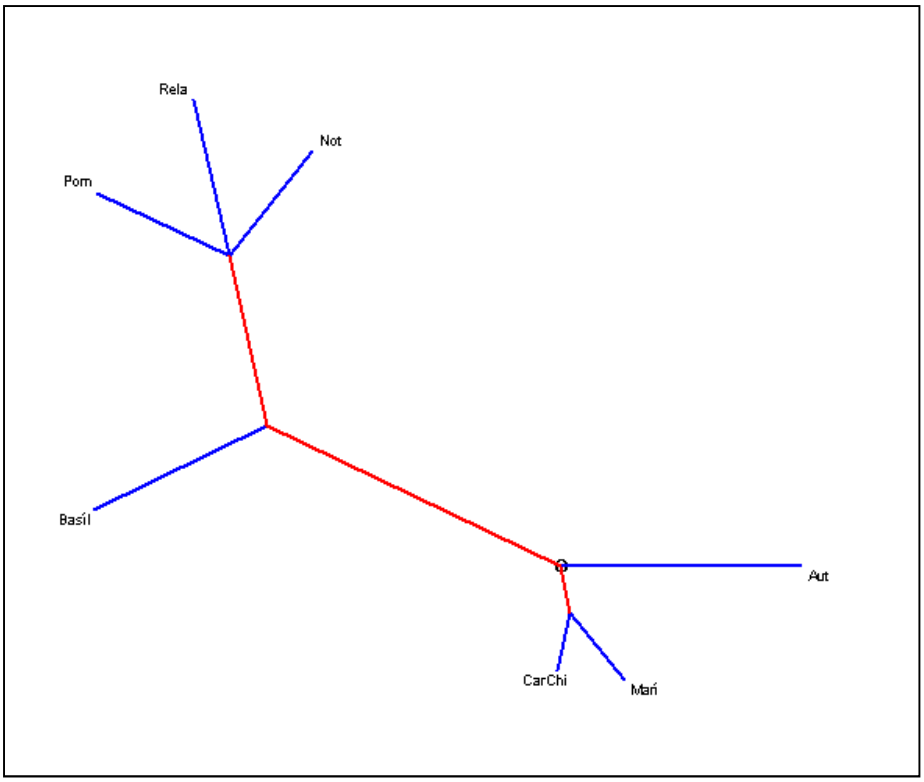

Gráfico 3 – Análise em árvore das altas frequências.

Pesquisadores como J. F. Burrows (1987; 1989) desenvolveram técnicas de atribuição de autoria baseadas especialmente nas frequências das formas encontradas nos textos. Sobre as técnicas de Burrows, em artigo intitulado *Computational Methods in Authorship Attribution* e publicado no *Journal of the American Society for Information Science and Technology* Moshe Koppel, Jonathan Schler e Shlomo Argamon argumentam que:

> The idea is to visualize the differences between texts written by different authors by projecting high-dimensional word-frequency vectors computed for those text onto the 2-dimensional subspace spanned by the two principal components; if good separation is seen between documents known to be written by different authors, then new texts may be attributed by seeing which authors' comparison documents are closest to them in this space (KOPPEL, SHLER e ARGAMON, 2009, p. 06).

Mesmo não dispondo das técnicas multidimensionais de Burrows, acreditamos que a análise feita pelo Hyperbase nos dá mais uma garantia de que a *Relação*

Abreviada e os textos de Pombal se aproximam, agora, no que tange às altas frequências.

Observando o gráfico, percebemos que além de estar ligada às obras do Marquês de Pombal, a *Relação Abreviada* também se liga às notas de *O Uraguai*. Um dos motivos está relacionado à igualdade do gênero entre os três textos. Outro motivo, talvez de maior peso, é a semelhança temática dos textos: uma crítica aos jesuítas.

Ainda no que diz respeito às altas frequências, podemos observar como as palavras funcionais se distribuem no *corpus*. Retiramos da tabela as palavras de sentido e deixamos apenas as funcionais.

Ordem	**Freq.**	**Forma**	28	1004	na	60	413	nos
3	9330	que	29	982	no	62	401	mas
4	7545	e	30	941	dos	63	397	pelo
5	6637	de	33	802	seu	65	386	seus
6	6575	o	35	792	mesmo	69	374	sem
7	5633	a	36	791	sua	72	338	nas
9	3917	se	38	682	à	75	335	desta
10	3023	do	40	629	como	76	334	tudo
11	2267	os	41	610	ou	78	309	assim
12	2073	em	44	559	já	79	308	onde
13	2024	não	45	553	das	82	297	alguma
14	1981	da	46	550	me	84	291	esta
16	1691	as	47	529	uma	87	287	todos
17	1639	lhe	48	487	este	88	286	sobre
18	1575	com	49	475	eu	90	280	logo
20	1452	por	52	457	também	91	279	nesta
21	1422	para	54	444	aos	94	271	aquele
22	1330	mais	55	442	nem	95	270	bem
23	1308	ele	56	434	quem	96	266	ainda
24	1302	ao	57	430	só	100	259	depois
25	1115	um	59	416	quando			

Tabela 5 – Palavras funcionais.

É interessante mencionar que mais recentemente importantes estudos sobre atribuição de autoria têm dados às palavras funcionais destacada função. Podemos, inclusive, encontrar trabalhos que utilizaram para a análise apenas as palavras funcionais. É o caso de: *Measuring the Usefulness Of Function*

Words for Authorship Attribution, de Shlomo Argamon e Shlomo Levitan (2005), do Ilinois Institute of Technology; *Effective and Scalable Authorship Attribution Using Function Words*, de Ying Zhao e Justin Zobel (2005), da Universidade de Melbourne e *Function Words in Authorship Attribution Studies*, de Antonio Miranda García e Javier Calle Martín (2005), da Universidade de Málaga. Como a bibliografia citada anteriormente nos mostra, as palavras funcionais podem e são utilizadas como medidas de estilo e consequentemente de atribuição de autoria.

Um dos primeiros trabalhos envolvendo as palavras funcionais e a atribuição de autoria foi desenvolvido por Mosteller e Wallace (1964), no estudo dos *Federalist Papers*. Os pesquisadores foram capazes de comprovar que uma pequena quantidade das palavras mais utilizadas em uma língua, as palavras funcionais, serviam como indicadores de estilo e de autoria. Mesmo após décadas e com ferramentas computacionais muito mais precisas, as palavras funcionais continuam sendo utilizadas como marcadoras de estilo. Para García e Martín (2005, p. 01): "Due to their high frequency in the language and highly grammaticalized roles, function words are very unlikely to be subject to conscious control by the author".

Se o que os pesquisadores supracitados dizem a respeito das palavras funcionais é verdade, isto é, que elas são utilizadas de forma inconsciente devido à alta frequência delas na língua, e que seu uso caracteriza o estilo de um autor, o cálculo do desvio das palavras funcionais em nosso corpus poderá nos ajudar a resolver nosso problema. Vejamos o gráfico!

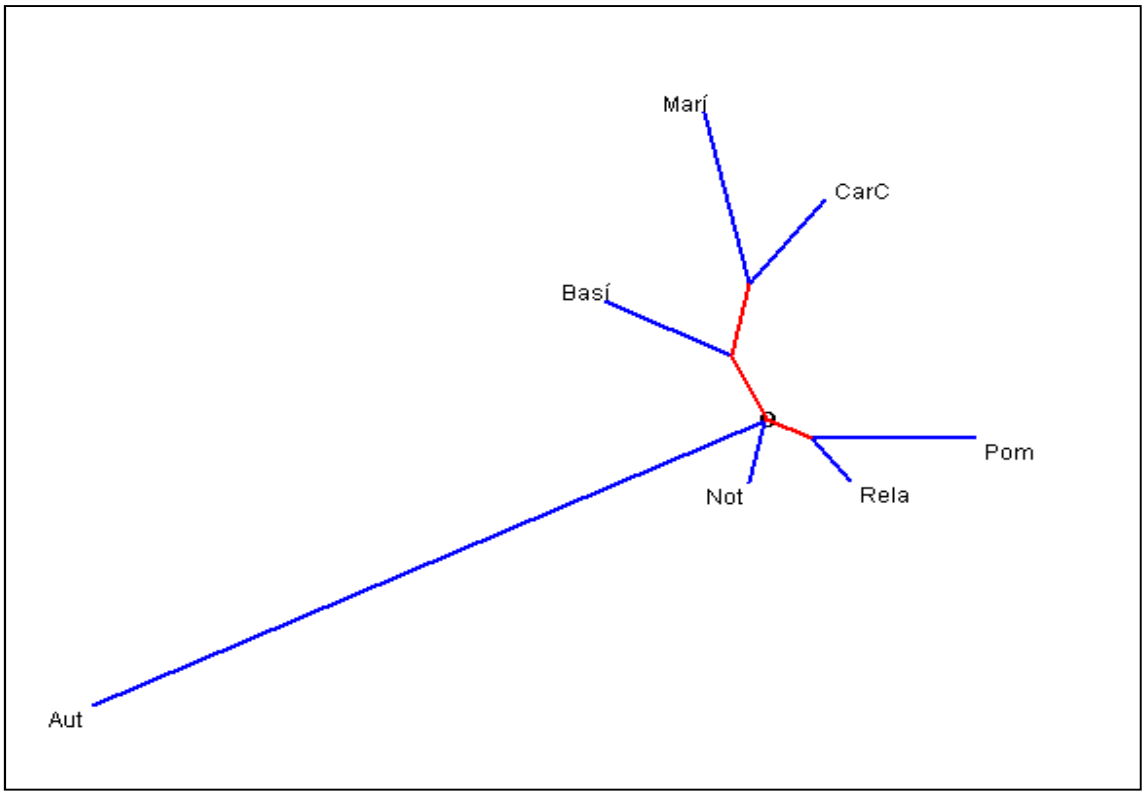

Gráfico 4 – Análise em árvore das palavras funcionais.

Novamente os textos de Pombal ligam-se à *Relação Abreviada*. Acreditamos que já seja possível, dada as aproximações de estilo entre as duas obras encontradas até então, começarmos a ter um posicionamento mais fundamentado no que tange à autoria da *Relação*. Mesmo assim, para nós, ainda é cedo para afirmar que o Marquês de Pombal escreveu a obra que estamos analisando. Mais análises precisam ser feitas.

Partindo agora para o campo dos códigos gramaticais, efetuaremos um cálculo a partir do que o programa chama de bicódigo, isto é, uma função do Hyperbase que mapeia a quantidade de vezes que dois códigos gramaticais são utilizados sequencialmente. Dessa forma, o sistema busca por frases que tenham um substantivo seguido de um verbo, um verbo seguido de um pronome, um adjetivo seguido de um determinante, e assim todas as possibilidades possíveis.

Ao mapearmos nosso *corpus*, encontramos 90 combinações de bicódigos diferentes. Utilizaremos o gráfico fatorial para que a visualização seja melhor.

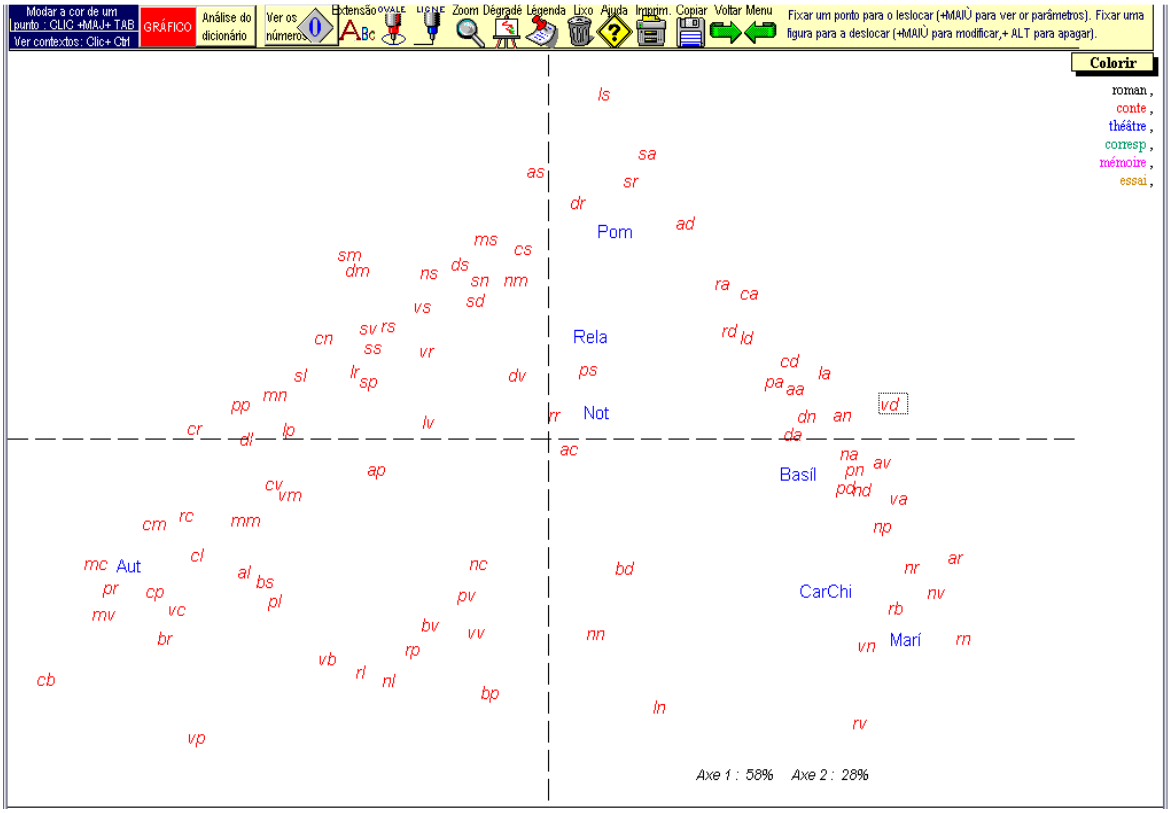

Gráfico 5 – Análise fatorial dos bicódigos.

Abaixo a lista dos códigos encontrados e suas siglas.

Código	Sigla
Substantivo	N
Verbo	V
Adjetivo	A

Numeral	M
Pronome	P
Advérbio	R

Determinante	D
Preposição	S
Conjunção	C

Tabela 6 – Códigos e suas siglas.

Na análise em árvore, onde as aproximações entre os textos são visualizadas de maneira mais privilegiada, as obras tiveram a disposição seguinte:

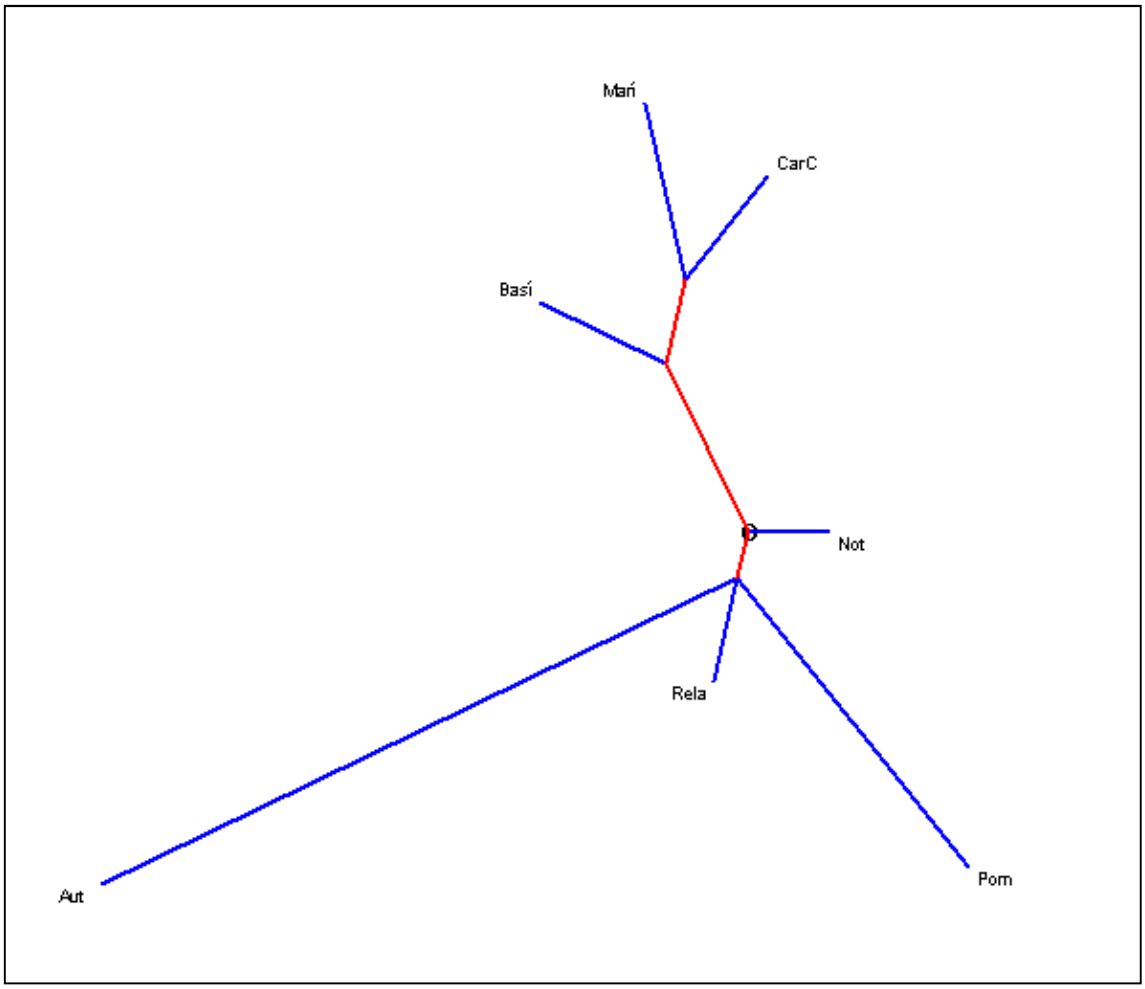

Gráfico 6 – Análise em árvore dos bicódigos.

Além da análise em bicódigos, também realizamos uma busca por triódigos. Quer dizer: as sequências possíveis do conjunto de três códigos. O arranjo dos textos se deu de maneira bastante parecida com a que constatamos na busca por bicódigos. Assim, mais uma vez a *Relação Abreviada* e os texto de Pombal se aproximam.

De maneira bastante interessante, durante as análises realizadas e nas visualizações dos gráficos, percebemos que as *Cartas Chilenas* se aproximavam de *Marília de Dirceu*, obra escrita por Tomás Antônio Gonzaga. Mesmo não tendo sua autoria materialmente comprovada, importantes estudos, como o de Afonso Arinos de Melo e Franco (1940) atribuem a autoria das *Cartas* a Tomás Antônio

Gonzaga, mesmo autor de *Marília*. Como nosso objetivo não é trabalhar com as *Cartas Chilenas*, deixaremos um análise de sua autoria para estudos posteriores.

Pelo que percebemos até agora, as semelhanças de estilo entre a *Relação* e os textos do Marquês de Pombal se dão em vários níveis: na pontuação; nas altas frequências; nas palavras funcionais; na sequência de dois códigos gramaticais e também na de três. De certa maneira, já podemos dizer que, se os textos não têm o mesmo autor, a semelhança de estilo entre os dois é de, literalmente, precisão matemática.

Tendo em vista as várias críticas que os métodos não tradicionais de atribuição de autoria sofreram e sofrem ainda dentro da academia, resolvemos aplicar os testes feitos em nossa primeira base textual a textos de autoria já conhecida. É de capital importância para nós que este estudo passe por diversos procedimentos até podermos, se possível, apontar um resultado final. Como já mencionamos desde o início do texto, nos é mais prudente e seguro, ainda, não levantar afirmações acabadas.

A Prova dos Nove

Para comprovar que alguns dos dados observados durante as análises realmente servem como parâmetro de estilo e de autoria, resolvemos criar uma base com diversos textos de autores já conhecidos e incluir, também, além da *Relação Abreviada* e dos textos do Marquês de Pombal, *O Uraguai,* para, assim, observar como o texto de Basílio da Gama se comportava em uma base maior. A base ficou da seguinte forma:

Obra	Ocorrências	Formas	Abreviação
Textos do Marquês de Pombal	29.786	5.165	Pomb
O Uraguai	10.611	2.704	Urag
O Cortiço	99.809	11.196	Cortiço
Casa de Pensão	118.542	12.303	CaPen

O Mulato	120.982	13.109	Mulato
São Bernardo	46.036	7.308	SBerna
Vidas Secas	32.051	5.356	ViSeca
A Moreninha	58.883	6.953	More
A Luneta Mágica	66.734	8.552	Luneta
Clara dos Anjos	62.542	8.593	Clar
Os Bruzundangas	49.712	8.642	Bruzun
O Moço Loiro	145.564	10.765	MLoiro
Dom Casmurro	79.292	8.632	DCasm
Brás Cubas	76.369	9.818	BCuba
Relação Abreviada	7.729	1.862	Rela

Tabela 7 – Textos da base 2 e o respectivo número de suas ocorrências e formas.

As obras foram agrupadas de acordo com sua autoria, excluindo dessa ordem os textos do Marquês de Pombal e a *Relação Abreviada*, uma vez que estes são os textos que estamos aproximando estatisticamente.

Nossa primeira análise foi quanto à pontuação dos textos. Tendo em vistas reduzir um pouco a quantidade de gráficos no nosso estudo, optamos por fazer uma análise que nos permitisse analisar mais de um fator ao mesmo tempo. Sabendo que o Hyperbase nos dá algumas possibilidades quanto a esse tipo de análise, resolvemos colocar em um mesmo gráfico, a pontuação e os bicódigos.

Novamente o sistema encontrou 90 possibilidades de bicódigos ao longo do *corpus*. Acrescentamos aos 90 bicódigos encontrados, os sete tipos de pontuação que utilizamos na nossa primeira base. Vejamos como os dados ficaram distribuídos em uma análise por fatores:

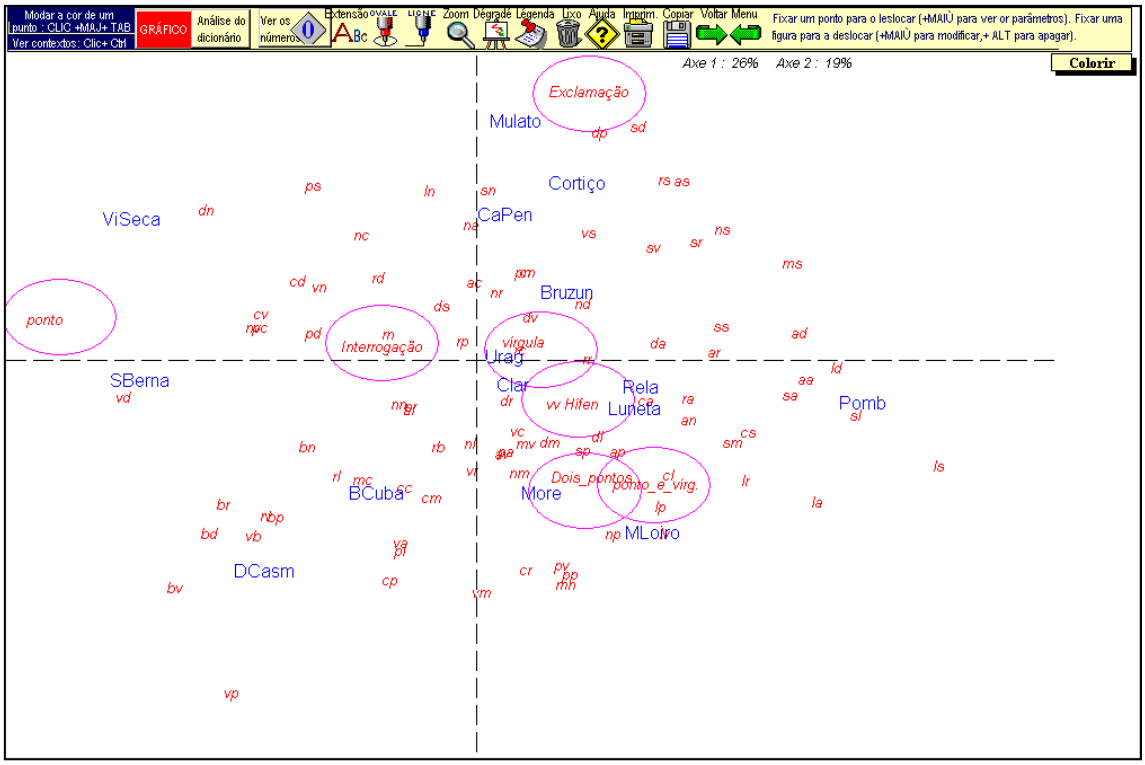

Gráfico 7 – Análise fatorial dos bicódigos e pontuação na base 2.

No gráfico acima, podemos perceber a disposição dos textos nos quadrantes e a aproximação entre eles a partir dos bicódigos e da pontuação. Numa rápida olhada, podemos perceber que as obras, em quase sua totalidade, estão dispostas nos quadrantes a partir de sua autoria. É o caso de *Brás Cubas* e *Dom Casmurro*, *O Mulato*, *O Cortiço* e *Casa de Pensão*, *A Moreninha*, *O Moço Loiro* e *A Luneta Mágica*. Em posições próximas também estão *Vidas Secas* e *São Bernardo*, bem como *Os Bruzundangas* e *Clara dos Anjos*. Podemos ver que *A Relação Abreviada* e os textos do Marquês de Pombal também estão em um mesmo quadrante, o que não acontece com *O Uraguai*.

Se colocarmos os mesmo dados, agora em análise em árvore, teremos o seguinte gráfico:

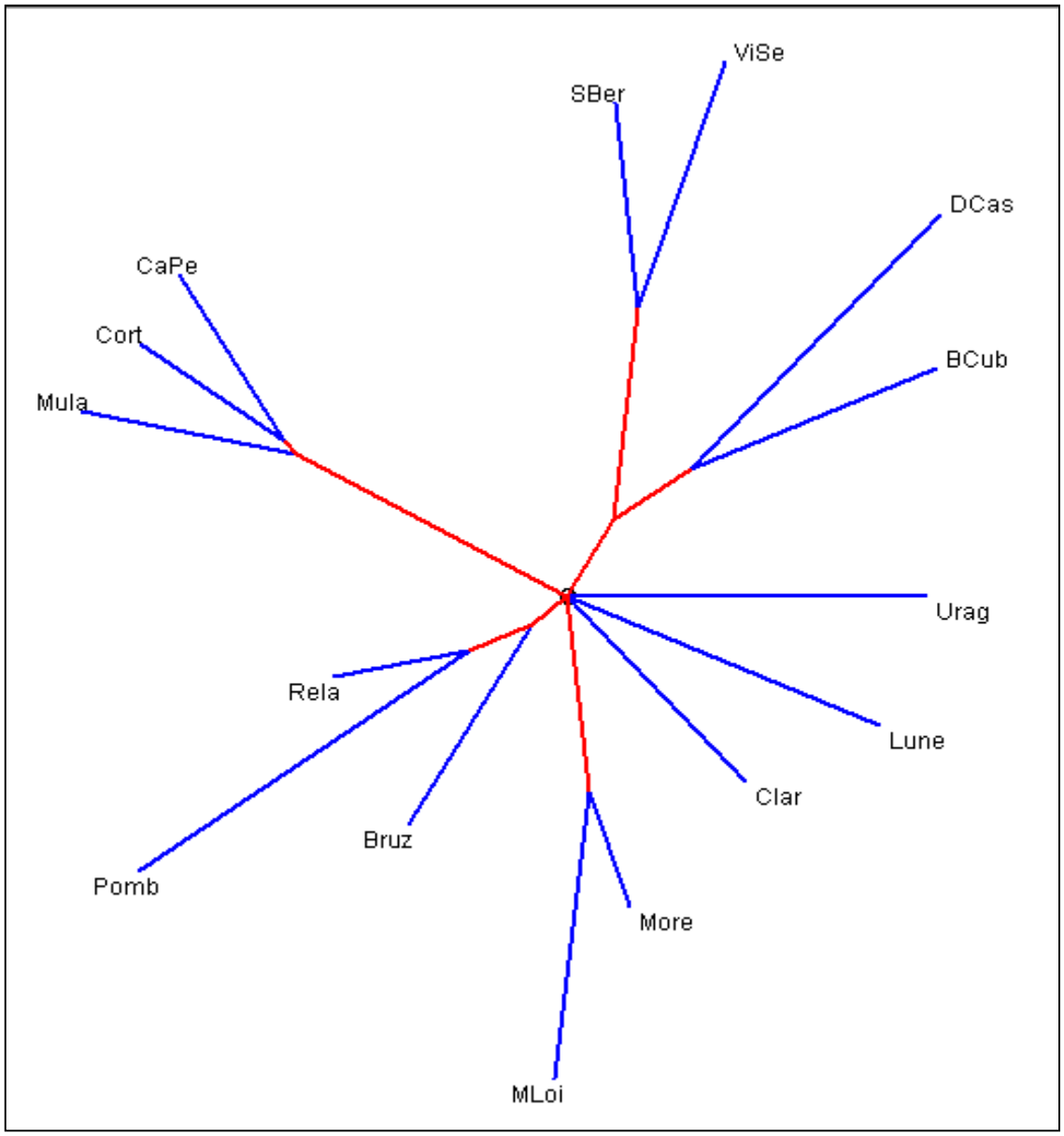

Gráfico 8 – Análise em árvore dos bicódigos e pontuação na base 2.

A partir da leitura dos dois gráficos, podemos ver claramente que os textos com o mesmo autor ou estão numa mesma raiz ou estão bem próximos. Assim, baseados na observação dos dados, é válido afirmar que os critérios iniciais que adotamos, a pontuação e a repetição dos bicódigos nos textos, servem como elementos marcadores de estilo e de autoria.

Dessa forma, observando os dados iniciais e como os textos com autores conhecidos se comportaram nessa segunda etapa, podemos afirmar com mais

segurança que há poucas chances de a *Relação Abreviada* ter sido escrita por Basílio da Gama.

Para, mais uma vez, validar o que os dados iniciais apontaram e demonstrar que os critérios escolhidos servem para os estudos de atribuição de autoria, aplicamos à segunda base, a que continha os textos de autores conhecidos, o cálculo do conjunto das frequências encontradas no *corpus*.

Podemos perceber que na segunda base os textos também se distribuíram no gráfico de acordo com a autoria. Colocados agora em um gráfico retangular, usando os mesmos procedimentos estatísticos utilizados nos gráficos de análise em árvore, vemos que além de se aproximar dos textos de Pombal, a *Relação Abreviada* também está no mesmo galho de *O Uraguai*, o que já havia acontecido no gráfico em árvore que usamos na primeira base. Como o gráfico é realizado a partir do conjunto das frequências das palavras, é esperável que as duas obras estejam próximas, uma vez que os textos são curtos e partilham da mesma temática.

Vejamos o gráfico:

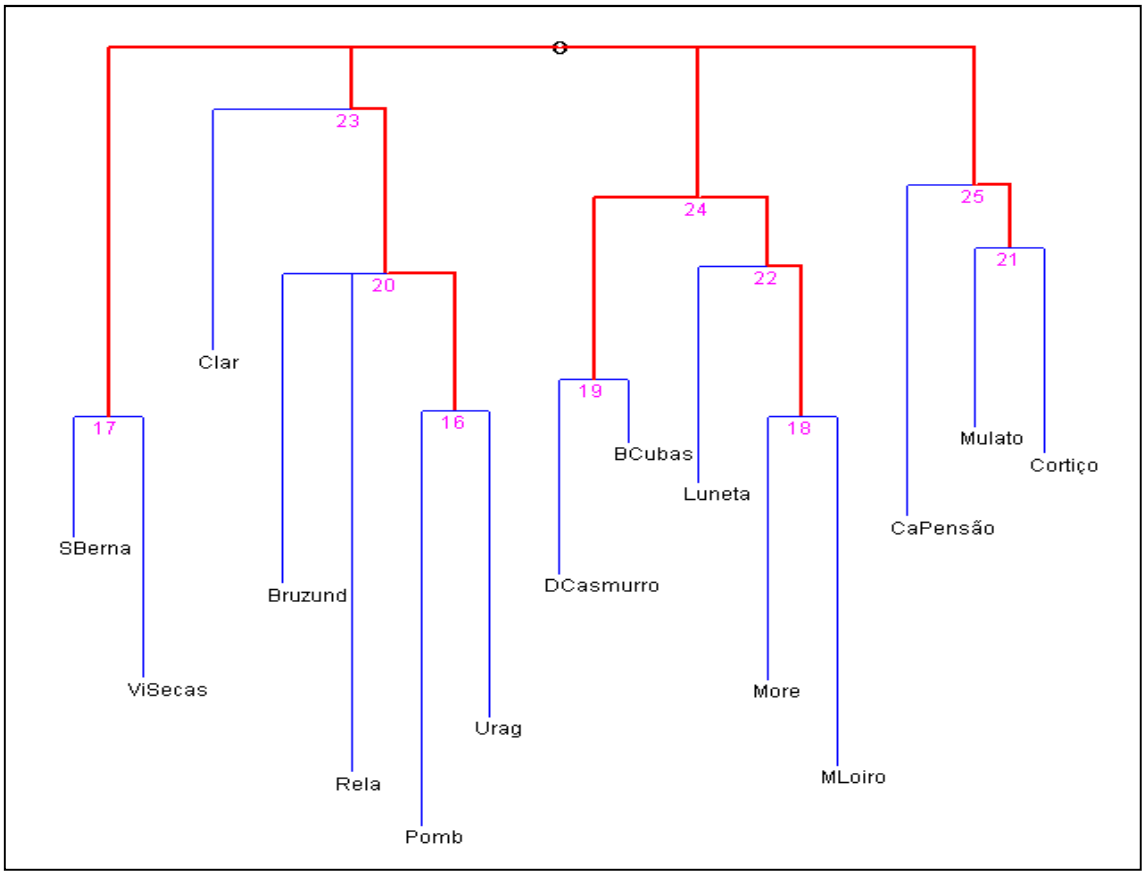

Gráfico 9 – Análise em árvore das altas frequências na base 2. Visualização retangular.

Mesmo sabendo dos motivos que faziam com que, na análise das frequências, a *Relação* se aproximasse de *O Uraguai*, resolvemos fazer um último teste também utilizando frequências. Com a ajuda do Hyberbase, buscamos os segmentos de palavras e/ou expressões que eram encontrados na *Relação Abreviada*, nos textos do Marquês e na epopeia de Basílio da Gama. Na primeira busca procuramos por segmentos que estavam presentes na *Relação* e nos textos do Marquês. Os segmentos foram os seguintes:

	Segmentos textuais encontrados na *Relação* e nos textos do Marquês de Pombal			
	Sequência	Total em todas as obras	Total nas obras relacionadas	Porcentual %
1.	ao tempo em que	9	3	33,33%
2.	do mesmo ano de	7	7	100,00%
3.	ao mesmo tempo	27	17	62,96%
4.	corte de lisboa	11	10	90,91%
5.	da mesma sorte	24	8	33,33%

6.	de sua majestade	32	32	100,00%
7.	do ano próximo	8	5	62,50%
8.	do capitão general	48	7	14,58%
9.	e muito menos	13	2	15,38%
10.	pela outra parte	8	8	100,00%
11.	por outra parte	10	10	100,00%
12.	por uma parte	9	9	100,00%
13.	que depois de	8	4	50,00%
14.	que sua majestade	14	14	100,00%
15.	capitão general	52	11	21,15%
16.	das referidas	12	8	66,67%
17.	dos portugueses	10	2	20,00%
18.	dos referidos	18	10	55,56%
19.	francisco xavier	11	6	54,55%
20.	grande número	10	10	100,00%
21.	nas referidas	16	3	18,75%
22.	pelo contrário	13	3	23,08%
23.	próximo passado	14	8	57,14%
24.	sua majestade	83	83	100,00%
25.	todos aqueles	9	8	88,89%

Tabela 8 – Segmentos textuais encontrados na *Relação Abreviada* e nos textos do Marquês de Pombal.

Colocados em um gráfico, os segmentos ficaram da seguinte maneira:

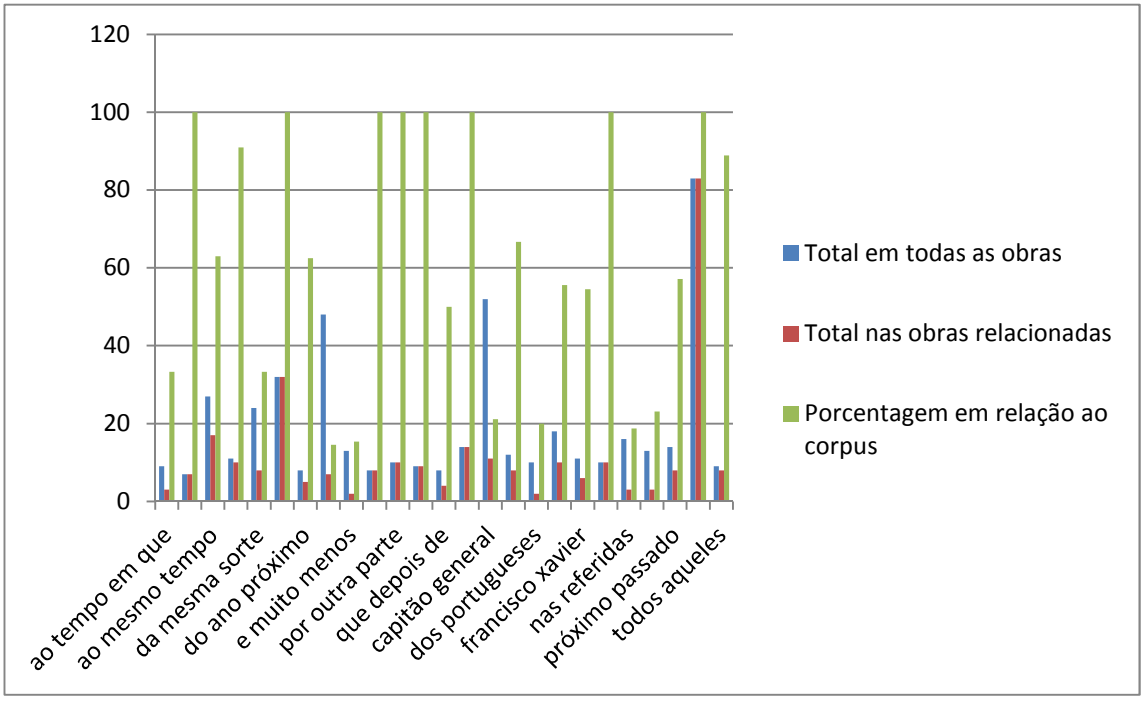

Gráfico 10 – Segmentos textuais encontrados na *Relação Abreviada* e nos textos do Marquês de Pombal.

Vinte e cinco segmentos foram encontrados. Os segmentos foram relacionados entre as duas obras destacadas e entre os outros textos do *corpus*, assim poderíamos ter uma ideia da frequência dos segmentos em relação às duas obras e em relação às outras. Por exemplo: *ao tempo em que* aparece 9 vezes em todo o *corpus* e 3 vezes nas obras selecionadas, dando um total de 33,3%; já o segmento *do mesmo ano de* está presente apenas nas obras relacionadas e aparece 7 vezes.

Como podemos notar, o número de segmentos que aparecem nas duas obras é relativamente alto. Mais interessante ainda, é perceber que alguns segmentos aparecem apenas nas obras relacionadas.

Os segmentos encontrados na *Relação* e no *O Uraguai* foram:

Segmentos textuais encontrados na *Relação* e no *O Uraguai*			
Expressões	Total em todas as obras	Total nas obras relacionadas	Percentual %
de outra sorte	14	3	21,43%
que os padres	9	5	55,56%
tinha mandado	9	2	22,22%

Tabela 9 - Segmentos textuais encontrados na *Relação Abreviada* e no *O Uraguai*.

É clara a redução dos segmentos encontrados. Dessa forma, mesmo havendo uma aproximação entre a *Relação* e *O Uraguai* no que diz respeito às altas frequências, quando analisamos as obras focando nos segmentos, a proximidade entre as obras não é significativa. Mais uma vez *A Relação* e os textos do Marquês de Pombal parecem ter sido escritos pela mesma pessoa.[8]

Conclusões

[8] Fizemos a busca por segmentos comuns presentes na *Relação Abreviada* e nas notas de *O Uraguai*, dessa maneira ficando com textos do mesmo gênero. O número de segmentos não superou uma dezena.

Nossa proposta inicial era fazer uma análise estatística do estilo da *Relação Abreviada* e, ao final, poder atribuir, com o mínimo de rigor técnico, um autor à obra.

A hipótese era que o Marquês de Pombal poderia ter escrito a *Relação*, já que o possível autor e a publicação da obra partilhavam de um contexto histórico e temporal semelhante.

Utilizando o Hyperbase, programa de tratamento estatístico de textos, colocamos nossa hipótese à prova e vimos que os resultados iniciais nos davam uma boa margem de proximidade entre *A Relação Abreviada* e documentos escritos pelo Marquês de Pombal.

Ao longo de todo o texto, mantivemos uma posição mais cautelosa antes de fazer qualquer tipo de afirmação. Como se trata de um estudo de atribuição de autoria, não podíamos correr o risco de fazer julgamentos precipitados. Por isso, quase que exaustivamente, fizemos os mais diversos tipos de testes possíveis quanto ao estilo da obra anônima. Foram testes relacionando a pontuação, os códigos gramaticais, as altas frequências, as palavras funcionais, os segmentos repetidos. Para não mencionar outros testes que não expomos aqui, mas que também nos davam dados que aproximavam as duas obras, como a análise das letras mais usadas e um mapeamento da estrutura sintática.

Após todos esses testes, podemos afirmar que, estatisticamente, o estilo empregado na *Relação Abreviada* e o estilo encontrado nos textos do Marquês de Pombal são mais que semelhantes. Acrescentando a esse dado o fato de que, segundo J. Lúcio de Azevedo em *O Marquês de Pombal e Sua Época* (1922), a *Relação Abreviada* foi redigida na Secretaria de Estado, e essa, à época, respondia ao Marquês, podemos acreditar que há envolvimento de autoria entre o texto e Pombal.

Como se trata de um estudo de menor fôlego, acreditamos que outras conclusões podem ser retiradas da aproximação entre os textos. Outras que possam fazer análises mais aprofundadas e baseadas em mais critérios. Enquanto não contamos com tais análises, continuaremos acreditando que o que foi demonstrado aqui seguiu padrões confiáveis e, assim, são válidos. De qualquer forma, cremos que esse trabalho possa servir para ao menos ser confrontado futuramente.

AUTHORSHIP ATTRIBUTION USING STATISTICAL ANALYSIS: AN EXPERIENCE WITH *RELAÇÃO ABREVIADA*

ABSTRACT: This article aims, from statistical and stylometric analysis and using the HyperBase, word-processing software that, by recognizing the structural peculiarities of the language, maps the textual categories and transforme them into statistical data, perform a rapprochement between the *Relação Abreviada*, text of unknown authorship, and their possible authors: o Marquês de Pombal and Basílio da Gama. Therefore, procedures and analyzes involving scores of texts, high frequency function words, repetition and other passages were performed and gave us heavy subsidies to believe that the *Relação Abreviada* may have been written by the Marquês de Pombal.

KEYWORDS: Authorship attribution. Statistics. Hyperbase. Marquês de Pombal. Basílio da Gama.

REFERÊNCIAS

ARGAMON, Shlomo; LEVITAN Shlomo. Measuring the Usefulness of Function Words for Authorship Attribution. **Proceedings of the 2005 ACH/ALLC Conference**, Victoria, 2005.

AZEVEDO, J. Lúcio de. **O Marquês de Pombal e Sua Época**. Rio de Janeiro: Annuário do Brasil, 1922.

BLAKE, Augusto Victorino Alves Sacramento. **Diccionario Bibliographico Brazileiro.** Rio de Janeiro: Imprensa Nacional, 1883. v. 4.

BRUNET, Etienne. **La punctuation et le rythme du discourse** (d'après les données du Trésor de la langue française), Nice, CUMFID, Université de Nice, n. 13, jul. 1981. p. 1-27. Disponível em : <http://literaturabrasileira.ufsc.br/>. Acesso em: 20 ago. 2012.

BURROWS, J. F. **Computation into Criticism**: A Study of Jane Austen's Novels and an Experiment in Method. Oxford: Clarendon Press, 1987.

_____, J.F. 'An ocean where each kind': Statistical analysis and some major determinants of literary style. **Computers and the Humanities,** v.23, n. 4, 1989. p. 309-321.

CHAVES, Vânia Pinheiro. **O Despertar do Gênio Brasileiro**: uma leitura de O Uraguai de José Basílio da Gama. Campinas: Editora da Unicamp, 2000.

FRANCO, Afonso Arinos de Melo. **Cartas Chilenas**. Rio de Janeiro: Imprensa Nacional, 1940.

FREITAS, Deise J. T. de. **A composição do estilo do contista Machado de Assis**. 204 f. Tese (Doutorado em Literatura) – Programa de Pós-Graduação em Literatura, Universidade Federal de Santa Catarina, Florianópolis, 2009.

GARCÍA, Antonio Miranda; MARTÍN, Javier Calle. The validity of lemma-based lexical richness in authorship attribution: A proposal for the Old English Gospels. **ICAME Journal**, n. 29, 2005. p. 115-129.

KOPPEL, M.; SCHLER, Jonathan; ARGAMON, Shlomo. Computational Methods in Authorship Attribution. **Journal of the American Society for Information, Science and Technology**, v. 60, n. 1, 2009. p. 9-26.

MELO, Sebastião J. de Carvalho e. **Cartas e Outras Obras Selectas do Marquês de Pombal**. Lisboa: C. Sanches, 1861. Disponível em: <http://archive.org/details/cartaseoutrasobr02pomb>. Acesso em: 05 ago. 2012.

MOSTELLER, F.; WALLACE D.L. **Inference and Disputed Authorship**: The Federalist. Addison Wesley, 1964.

ZHAO, Ying; ZOBEL, Justin. Effective and Scalable Authorship Attribution Using Function Words. In: **Information Retrieval Technology**. Jeju Island, 2005.

Texto enviado em maio de 2013.

Texto aprovado em julho de 2013.