Em Busca da "Impressão Digital" de um Texto

Saulo Cunha de Serpa Brandão

Departamento de Letras – Universidade Federal do Piauí (UFPI)
64.055-450 – Teresina – PI – Brasil

saulo@teacher.com

Resumo: *Esse escrito descreve uma pesquisa em seu estágio inicial. A busca é por um método para definir autoria de textos apócrifos, a partir de abordagens quantitativa e qualitativa dos mesmos e com a utilização um software chamado LEXICO3 (CLA2T/Paris 3).*

Abstract: *This paper describes a research in its initial stage. The search is for a method to define authorship of apocrypha texts, departing from quantitative and qualitative approaches of them, and using a software named LEXICO3 (CLA2T/Paris 3).*

1. Introdução

Esta pesquisa encontra-se em sua fase embrionária. O vetor que nos leva a apresentar este "paper" é muito mais para tentar encontrar parceiros para um diálogo construtivo e lançar as primeiras bases metodológicas que guiarão o trabalho, do que para apresentar resultados palpáveis. Estes serão amostras mínimas e, provavelmente, temporárias do que resgatamos nas primeiras simulações.

2. Do Objetivo e Objeto

O objetivo primeiro da pesquisa é verificar a possibilidade de se obter vestígios quantitativos e/ou qualitativos na escrita que possam servir para evidenciar a autoria de textos apócrifos. Ou seja, determinar se a pessoa, ao escrever, deixa algum tipo de índice que singularize sua escrita. O primeiro objeto escolhido, *per se*, demonstra que profundidade e seriedade desejamos impor a este projeto, trata-se de *Cartas Chilenas*. Elas são um conjunto de 14 poesias satíricas que datam do século XVIII e são atribuídas a Tomás Antônio Gonzaga, mas foram assinadas com um pseudônimo: Critilo. Mas existe muito para ser desvendado desse mistério: primeiro, as *Cartas* foram encontradas em dois manuscritos diferentes, um contém as sete primeiras cartas e outro as 7 complementares; segundo, existem quatro manuscritos, ligeiramente diferentes, da mesma época, contendo as sete primeiras cartas; terceiro, a 14ª carta é atribuída a Cláudio Manuel da Costa.

Em uma revisão bibliográfica vamos encontrar toda as crítica e história da literatura do século XX apontando Gonzaga como autor das 13 cartas (1 a 13), de Cláudio seria a 14ª. Já a revisão do século XIX mostra vozes discordantes, como a de Silvio Romero (1980, p.429) que indica Alvarenga Peixoto como provável autor de *Cartas*. O dito popular sobre opinião unânime, nos impulsiona a ir fundo nesta investigação e dela tirar lições que nos permitam criar um mecanismo para averiguação de qualquer texto apócrifo, ou de origem conhecida mas de competência questionada.

A diferença das análises feitas no século passado e século XIX para a que pretendemos empreender agora, sobre esse objeto, está nas ferramentas disponíveis. Tudo que levantamos sobre as tentativas de definição da autoria de *Cartas* está baseado em estudos feitos sobre dados estilísticos (qualitativos). São investigações sobre figuras de imagem ou de discurso, versificação, qualidade das rimas. Uns poucos, como Varnhagem (1854 [apud Oliveira, 1972]), apreciam as qualidades semânticas do léxico utilizado. Mesmo estes, limitam-se à apreciação de uma dúzia de palavras, não mais.

3. O Software

Para a empreitada que ora propomos, utilizaremos um software desenvolvido na Université de la Sorbonne Nouvelle – Paris 3, pela equipe CLA2T, denominado LEXICO3. Este é um programa de aplicação lexicométrica extremamente versátil e de utilização não muito complexa -- vale lembrar nossos leitores que somos especialistas em literatura e que a nossa lida com elementos algorítmicos não é tão amigável.

O LEXICO3 tem se mostrado uma ferramenta muito poderosa e de simples utilização. Ele nos permite, de forma ágil, balizar livremente o texto a ser analisado, determinando como dividir o texto, fazer contagens dentro de um balizamento, determinar o tamanho de um segmento repetido a ser pesquisado, fazer o levantamento das ocorrências do segmento repetido, indicar a distribuição das palavras dentro do texto, expor as concordâncias que ocorreram com uma palavra, elaborar gráficos indicando as freqüências relativa e absoluta da aparição de uma palavra em uma determinada baliza, etc.

Isso não quer dizer que não seja trabalhoso lidar com esse tipo de análise. O texto tem que estar completamente digitalizado e as balizas distribuídas obedecendo uma seqüência e lógica próprias do software que tomam muito tempo para serem entendidas e executadas. Um balizamento errado leva a uma "janela" informando muito pouco do erro cometido – esta, talvez, seja uma das maiores fraquezas do software -- o que leva o pesquisador a rever todo o balizamento do texto. Outro componente do programa que deixa muito a desejar é o manual de utilização.

4. As Frentes de Trabalho

Outro tipo de ordenamento que se faz necessário, diz respeito aos tipos de análises que faremos. Estamos trabalhando em duas frentes: análise quantitativa e qualitativa. Este embrião que ora apresento é o começo do desenvolvimento da parte quantitativa. Para iniciar esse trabalho tivemos que determinar quais seriam os elementos, ou formas, que seriam investigados, algumas possibilidades que se apresentaram eram as seguintes: tamanho das frases/palavras, riqueza do léxico, freqüência das palavras, uso da pontuação, freqüência de sinais de pontuação. As possibilidades são inúmeras.

Decidimos por visitar outros estudos desenvolvidos no passado para evitar começar o trabalho repetindo erros anteriores. Essa decisão foi muito apropriada e aprendemos muito com o que encontramos. T. C. Mendenhall (1901, [apud Peng e Hengartner, 2001]) teria feito gráficos com a freqüência que palavras longas apareciam nos escritos de Shakespeare e Bacon. Em 1975, C. B. Williams (idem) refez as "curves" de Mendenhall para descobrir que não existia evidência para comprovar qualquer das possibilidades. O mesmo Williams tinha feito um trabalho em 1940 tentando determinar

autoria usando o número de palavras por sentenças. Outros pesquisadores usaram esses mesmos parâmetros em outros corpora, mas os resultados não parecem animadores.

5. Os Primeiros Parâmetros

Nossa primeira tentativa está sendo com a contagem e distribuição de palavras funcionais (conjunções, dêiticos, preposições). A lógica na escolha dessas palavras é que elas são usadas de forma mais independente. As palavras funcionais são utilizadas de acordo com a necessidade de coerência textual e escolhida, geralmente, dentre um número reduzido de possibilidades. Essa escolha está justificada pelos resultados obtidos por Peng e Hengartner (2001) na análise procedida sobre textos de Austen, Carther, Doyle, Dickens, Kipling, London, Marlowe, Milton e Shakespeare.

Cumpre-nos informar, ainda, como foi dividido o objeto, *Cartas Chilenas*, para efeito de checagem de resultados. Partimos do princípio que não aceitaríamos as autorias indicadas na bibliografia compulsada, para fins metodológicos, e deixamos em aberto a indicação, dependendo dos achados da investigação. Como dissemos, as cartas estão divididas em dois manuscritos: um contendo as 7 primeiras cartas e o segundo com as demais. O passo posterior será determinar se a pessoa que escreveu o primeiro grupo de poesia foi a mesma que escreveu o segundo. A seguir, cotejar os achados nas diversas cartas com aqueles da 14ª. Após esse apanhado inicial, passaremos a comparar os achados em *Cartas Chilenas* com a produção de origem segura de poetas contemporâneos às *Cartas*. Neste primeiro momento estaremos cotejando as contagens encontradas para as cartas 2ª, 10ª e 14ª .

6. Os Primeiros Achados

No exemplo citado, acima, desenvolvido por Peng e Hengartner (2001), foram utilizadas 69 palavras funcionais na seleção das que melhor se prestavam para a averiguação de autoria. Em nossa pesquisa iniciamos por determinar que as palavras funcionais usadas em maior número seriam as primeiras a serem testadas, dessa forma começamos com: que, o, e, a, não, os, de, se, ao, um, aos, em, as, do, já, com, seu, da e do. As contagens que apresentaram relevo foram:

Palavra	**2ª**	**10ª**	**14ª**	**Palavra**	**2ª**	**10ª**	**14ª**
O	**54**	**59**	**41**	*As*	*13*	*20*	*16*
A	*31*	*43*	*46*	**Do**	**11**	**16**	**21**
Não	**26**	**24**	**12**	**Já**	**11**	**4**	**1**
Os	**21**	**23**	**32**	**Com**	**10**	**8**	**3**
Se	*20*	*10*	*19*	**Seu**	**10**	**9**	**1**
Um	**15**	**19**	**8**	**Da**	**6**	**6**	**23**
Aos	*14*	*6*	*6*				

Observa-se que O, NÃO, OS, UM, DO, JÁ, COM, SEU e DA apresentam um padrão que sugere estilos parecidos entre as cartas 2ª e 10ª e divergente em relação a 14ª. Já os léxicos em itálico remeteriam para outro tipo de interpretação.

A surpresa maior nessas primeiras simulações apareceram quando balizamos o LEXICO3 para que ele fizesse a contagem de segmentos repetidos em cada grupo de 10 formas diferentes. O intrigante começou com o número de pares encontrados em cada uma das cartas: na 2ª, 39 pares, na 10ª, 42 pares e na 14ª, 23 pares. Ou seja, o estilo do

autor da 14ª carta é muito econômico quando se trata de segmento repetidos, quase metade do encontrado nas outras duas.

Mais, o inesperado não parou por aí. Levantamos que 4 pares de palavras ocorrem com relativa alta freqüência na 2ª e 10ª cartas, e não se fazem presentes, sequer, uma vez na 14ª carta, são eles:

Segmentos	**2ª**	**10ª**	**14ª**
em que	3	4	O
o que	2	5	O
o nosso	5	4	O
que não	3	6	O

O pouco que coletamos até o presente nos estimula a continuar com a investigação. Verificamos que existem padrões quantificáveis, mais ou menos, fixos nos estilos dos autores das poesia em questão e que eles poderão se revelar a partir de estudos quantitativos. Existem ainda várias ferramentas no software escolhido que não aplicamos até agora.

7. Conclusão

O passo mais imediato da pesquisa, a seguir, é alargar as contagens para as outras cartas para checar se o padrão se define melhor e determinar um padrão mínimo que represente cada grupo de cartas. Para depois, começarmos a checar quais poetas contemporâneos às *Cartas* apresentam padrões semelhantes. Paralelamente a esta análise quantitativa, realizaremos um estudo qualitativo das poesias, na busca por um padrão estético que esteja mais presente nas *Cartas* e em outros escritos de autores para serem cotejados e que venha corroborar os achados quantitativos, ou que os rejeitem.

O domínio de uma técnica de definição de autoria que seja confiável trará incontestáveis avanços para estudos da genética textual e clareará cantos escuros da historiografia literária brasileira e outras. São inúmeros os escritos apócrifos em nossa literatura e esse desconhecimento faz com que pesquisadores enveredem por caminhos tortuosos e árduos com o objetivo de comprovar uma autoria, para depois, terem suas hipóteses fragilizadas pela quantidade de índices que se consegue colher manualmente.

8. Bibliografia

ÁVILA, Affonso.(1980) "O lúdico e as projeções do mundo barroco". 2ª ed. São Paulo: Perspectiva.

FERREIRA, Delson.(1986) "Cartas chilenas:retrato de uma época". 2ª ed. Belo Horizonte: Ed. UFMG.

GONZAGA, Tomás.(2003) "Cartas chilenas". Em **www.cce.ufsc.br/~nupill**, Abril.

OLIVEIRA, Tarquínio.(1972) "As cartas chilenas: fontes textuais". São Paulo: Editora Referência.

PENG, R. e Nicolas HENGARTNER.(2001) "Quantitative analysis of literary styles". The American Statistician. V.56, N 3, p. 175-185.

ROMERO, Silvio.(1980) "História da literatura brasileira". 7ª Ed., Rio de Janeiro: José Olympio, tomo 2. 5 tomos.