<REVISTA TEXTO DIGITAL>

ISSN 1807-9288
- ano 2 n.1 2006 –
http://www.textodigital.ufsc.br

ATRIBUIÇÃO DE AUTORIA

um problema antigo, novas ferramentas

AUTHORSHIP ATTRIBUTION

an old problem, new tools

Saulo Cunha de Serpa Brandão
Doutor em Letras pela UFPE
Universidade Federal do Piauí
brandaosaulo@yahoo.com

***RESUMO*:** Nesta pesquisa investigamos a eficiência da utilização de softwares especialmente desenvolvidos para fazer análises lexicométrica e estilométrica para determinação de autoria de textos anônimos. O programa utilizado foi o LEXICO3 e o objeto analisado foi Cartas Chilenas. O propósito inicial era utilizar o método proposto por Peng e Hangartner, contabilizando o número de vezes que léxicos funcionais apareciam em cada poema e comparar com os mesmos dados nas poesias assinadas pelos possíveis autores dos versos satíricos. O resultado da aplicação do método se mostrou instável e não conclusivo, mas outros números apareceram com padrão mais estável e que sugere que Cartas Chilenas foi escrito por muitas mãos o que torna a identificação dos autores uma tarefa muito difícil, senão impossível.

PALAVRAS-CHAVE: Cartas Chilenas. Atribuição de autoria. Análise lexicométrica. LEXICO3.

***ABSTRACT*:** In this research we investigate the efficiency of the use of software specially developed to proceed lexicographic and stylographic analyses to determine authorship of texts written by unknown authors. The software used was the LEXICO3 and the object of the research was Cartas Chilenas. At first, the purpose was to use the method proposed by Peng and Hangartner, counting the number of functional lexicons that appeared in each poem and compare the numbers with those collected from poems that were signed by the possible authors of the satiric verses. The result turned out to be unstable and not conclusive, but some other numbers appeared showing a pattern much more stable and that suggested that Cartas Chilenas was written by many hands what makes the task of identification very difficult if not impossible.

KEYWORDS: Cartas Chilenas. Authorship attribution. Lexicometric analysis. LEXICO3.

PARTE I - PREÂMBULO

Determinar a autoria de textos anônimos não é atividade nova no meio acadêmico e na tradição ocidental. Um exemplo didático e fecundo da antiguidade dessa atividade está primordialmente nos estudos dirigidos no meio religioso, com o objetivo de determinar a autenticidade de textos da tradição cristã, para afirmar o valor da reprodução dos ensinamentos e feitos de Jesus Cristo enunciados pelos apóstolos. O trabalho da determinação da autoria e, portanto, da autoridade desses textos para a fundação da tradição cristã, mobilizou estudiosos que, desde então e a cada época, retomam o debate do tema com instrumentos de trabalhos novos em função do objetivo de consolidar o conjunto de princípios que servem de base para o sistema religioso que pretende se estabelecer doutrinariamente. Esse trabalho de verificação da autoria, em função da autoridade a ser atribuída ao texto, estabelece, além disso, a diferença entre textos canônicos e os apócrifos numa determinada tradição religiosa. Tal trabalho é aprofundado continuamente, mesmo relativamente a documentos mais antigos e com autoridade reconhecida, como, por exemplo, quanto aos *Salmos*, para iluminar a pesquisa sobre as diferentes tradições teológicas que tais textos partilham com os demais do Antigo Testamento, esclarecendo as filiações e as disputas que mobilizavam os projetos teológicos de diferentes épocas. Nessa área específica, por exemplo, tal pesquisa ganhou sistematicidade no século XIX, fazendo com que a exegese demonstrasse que o livro vétero-testamentário[1] atribuído ao profeta Isaías é, de fato, trabalho realizado por três autores diferentes, abarcando três momentos históricos distintos, agrupado o material textual de modo a fortalecer um determinado projeto teológico. Este é só um exemplo que indica para o fato de que a determinação de autoria é ocupação importante de filólogos e teólogos, a mesma que passou a preocupar, mais recentemente, também os teóricos e críticos literários.

A prática de transcrição de textos, cuja autoridade deveria ser reconhecida, foi parte da atividade monacal também no período medieval, e a verificação de autoria de textos, desde então, está enraizada nos hábitos intelectuais europeus. Existe lá um permanente e fecundo exercício na elaboração de técnicas e métodos para determinação de

autoria de textos e esta preocupação chegou, por extensão, aos Estados Unidos da América. Ali, por exemplo, uma organização das mais prestigiosas no mundo das letras como a *Modern Language Association*, dedica, eventualmente, números temáticos de sua revista *Publications of the Modern Language Association* (PMLA) a trabalhos de verificação de autoria. No Brasil, desconhecemos trabalhos sistemáticos que tenham sido realizados nessa linha, mesmo que, por questões específicas de discussão de determinado texto participar do cânone literário autorizado, temos notícias de pesquisadores que se debruçam sobre um ou outro trabalho anônimo em busca da determinação de seu real autor, mas, até onde sabemos, essa atividade não resultou na elaboração de instrumentos consistentes de investigação. Essa questão fica muito clara quando nos utilizamos do *Google* para buscar informações de registros quanto ao termo "atribuição de autoria", ou procurando por expressões similares. O resultado de tais pesquisas em língua portuguesa indica menos de uma dúzia de sítios editados na internet com esses temas, enquanto, o mesmo levantamento com expressões similares em inglês ou em francês apresenta, como resultado, mais de seis mil citações.

A novidade do desenvolvimento de ferramentas telemáticas para tratamento de dados em grande volume, como a de busca, por exemplo, passou a oferecer outras possibilidades de trabalho, específicas e novas, agora também direcionadas para a área de determinação de autoria de textos. Um novo movimento de busca por esclarecimento quanto à autoria de textos discutidos pela tradição pode ser empreendido, agora fiado na possibilidade de levantamento de características de corpora nunca pensada anteriormente.

Como em outros casos de desenvolvimento de tecnologias e instrumental de trabalho, na contemporaneidade, esses avanços aconteceram inicialmente com intuitos pouco nobres. O Departamento de Defesa dos EUA, por exemplo, investiu pesadamente em softwares denominados genericamente de *Word Miners* (cf. Love, 2002). No início, estes foram utilizados para a procura por palavras indicadoras de ações subversivas em um número imenso de corpora, como: Cuba, bomba, explosivo, míssil. Os algoritmos primeiros se revelaram frustrantes pois apontavam para um arsenal de textos que delatavam periculosidades imediatas, quando na verdade eram, por

exemplo, listas de festas onde *Cuba-Libre* é uma bebida e *bomba*, um tipo de doce. Mas as ferramentas foram se sofisticando e incorporadas por outras áreas do conhecimento. E, como não poderia deixar de ser, a lingüística e a literatura vêm se apoderando delas para seus fins. Os lingüistas são mais organizados e têm pretensões mais cientificistas, assim criaram a, hoje, bastante difundida subárea da lingüística computacional. A literatura vem dando seus passos de forma mais cautelosa. Mas alguns, mais atrevidos, como o Prof. Donald Foster, já se consideram pesquisadores das áreas de *literary forensic* ou *computational literary*.

No cenário internacional existem números que justificam a pretensão de Foster, por exemplo: O'Donnell (1970) estima que nos cinqüenta anos anteriores aos da edição de seu estudo, cerca de trezentos livros e mais de três mil artigos tenham sido publicados sobre o assunto nos EUA e Europa. Na década de 70 do século passado, com o desenvolvimento e usos cada vez mais ordinários de computadores, um dos futuríveis apontava para a possibilidade de que em menos de trinta anos o problema de definição de autoria poderia estar completamente resolvido. Além disso, indicava que métodos de determinação de autoria estariam elaborados para lidar com os mais comuns entraves encontrados pelos estudiosos, como, por exemplo, o da identificação de textos de autoria múltipla. Essa fatura continua em aberto, uma vez que não existe um método definitivo para determinação de autoria e que, já há algum tempo, novas classes de especialistas estão se envolvendo com o problema, entre esses os psicólogos, os programadores e os estatísticos. Não existe, nessa área de trabalho acadêmico, garantia de que qualquer grupo esteja perto de encontrar uma fórmula definitiva.

A condição desses estudos estava meio estacionada, com os mesmos pesquisadores buscando técnicas apropriadas para a definição de autoria e testando novas ferramentas para aprimorar antigas fórmulas, ou inventando fórmulas novas, atividades possíveis graças ao desenvolvimento de novas tecnologias. Em 1996, o *New York Times* cedeu o milionário espaço de sua primeira página para noticiar a descoberta da autoria de uma elegia de 1612, de pouco mais de 578 versos,

como sendo de William Shakespeare, pelo professor Donald Foster. A peça em questão era *A funeral elegy for master William Peter* assinada pelas iniciais de um nome: W. S. Essa história é longa e não temos o espaço necessário neste *paper* para a descrição do trabalho realizado, mas é importante informar que o livro em que Foster indicava a autoria tinha sido publicado em 1989, resultado da tese doutoral do autor.

As iniciais W. S., a época em que se deu a morte em questão e a da elaboração da elegia, o fato de William Peter, o morto enaltecido, ter sido uma pessoa das relações de amizade do bardo inglês e outros indícios apontavam, com muita força, para Shakespeare como o autor do texto. Os peritos na dicção do poeta, porém, se negavam a aceitar a indicação pelo fato de a elegia ser um poema fraquíssimo, não condizente com a produção textual fenomenal de Shakespeare. Don Foster, utilizando uma perspicácia ímpar e, além disso, muito trabalho de gabinete e de campo, conseguiu demonstrar, de forma positiva e quase inquestionável, na época, a sua proposta de atribuição de autoria do texto estudado. Acima de tudo, para secundar a proposta construída, ele contava com uma erudição rara e foi auxiliado por uma memória privilegiada. Foster, apesar das evidências então indicadas, já reconhece hoje que a indicação de Shakespeare como autor da elegia estudada para definição da autoria não se sustenta, e as suspeitas que fundamentam novas pesquisas já indicam outros autores possíveis.

Na mesma semana em que o trabalho de Foster foi objeto de notícia da primeira página do *New York Times*, ele entrou em outra disputa de grande repercussão pública ao determinar o autor de *Primary Colors* como sendo o jornalista Joe Klein. Este passou vários meses negando a autoria, mas depois reconheceu ser o autor daquele romance. Todo esse redemoinho provocado em torno do debate sobre o tema da definição de autoria de textos publicados de modo anônimo ou sob pseudônimos deu vida nova a esse viés de pesquisa e chamou a atenção de especialistas de outras áreas para o desenvolvimento de ferramentas utilizadas nesses casos. Como parte da repercussão do seu trabalho, o próprio Foster se viu envolvido em outros casos de trabalhos de verificação de autoria, por exemplo os da área da criminalística que envolviam confirmação, para fins forenses, da similaridade

dos estilos de redação de texto, tanto no caso que envolvia Ted Kaczynski e o terrorista norte-americano Unabomber; naquele da disputa Lewinsky-Tripp, como parte do processo político que estudava o envolvimento do presidente estadounidense, Clinton, com uma estagiária da Casa Branca, para destacar dois dos mais famosos e menos sigilosos.

Don Foster, com seus acertos e erros, contribuiu para reaquecer o interesse pelos estudos de determinação de autoria e em seus dois livros sobre o assunto apontou um caminho profícuo para os estudos na área, que ainda, como dissemos, depende muito de características pessoais do pesquisador, como: erudição, perspicácia, curiosidade e boa memória. Desses traços de habilidades pessoais, alguns podem ser desenvolvidos por pesquisadores interessados nesse tipo de atividade, mas outros são dotes que não dependem somente de nossa vontade ou determinação. Se pretendemos trabalhar sobre o tema da determinação de autoria de textos editados de modo anônimo ou sob pseudônimo com base em um método científico, essa atividade não pode estar sustentada na aplicação de dons particulares, mas, sim, estar estabelecida sobre um conjunto de regras bem definidas e enunciadas objetivamente.

Love, Foster, O´Donnell são alguns dos estudiosos do tema que lançaram propostas para uma possível sistematização dos estudos de atribuição de autoria dos textos que estamos estudando. Eles concordam que existem dois tipos de índices que devem ser observados para a fundamentação de investigações nessa área de pesquisa; esses índices são formados pelas características de natureza intrínseca do texto e de natureza extrínseca. Os índices intrínsecos são aquelas pistas deixadas, conscientemente ou não, pelos autores na própria tessitura textual, enquanto os extrínsecos são aqueles dados levantados pelo pesquisador a partir do contexto sócio-histórico em que a obra apareceu. Um exemplo de dado levado em consideração para o exame do contexto social e histórico da produção de um texto é aquele da verificação do material textual produzido por outros autores no mesmo espaço temporal da aparição do texto em estudo, ou, ainda, aquele formado pelo exame da comunhão espacial entre os possíveis autores e o texto estudado, em

especial uma possível confissão de autoria feita pelo autor em outro texto.

Os teóricos atuais também são unânimes em determinar que, embora um índice extrínseco forte seja indispensável para a indicação de um autor, é no trabalho de análise textual que a determinação de autoria deve estar embasada. Além disso, concordam que a expectativa de formulação de uma técnica definitiva para a determinação de autoria de um texto ainda seja uma miragem metodológica. Por outro lado, os estatísticos que se preocupam com o assunto desarrefecem o ânimo dos que se aventuram nessa área de estudo, primeiro indicando que os resultados de análises textuais com base nessa ciência possam dar o ponto final para querelas autorais, isto porque eles nos ensinam que nenhum estudo estatístico do gênero poderia chegar à certeza absoluta, alcançando cem por cento de confiança no resultado, ao informar que um determinado poema foi escrito por um certo poeta. Haverá sempre um gradiente, pequeno que seja, entre a dúvida e a evidência.

Os índices extrínsecos (Love, 2002, p. 51) são provenientes da indicação dos dados relativos ao mundo social no qual o texto ficcional foi criado, promulgado e lido, isolado ou conjuntamente. Tais índices externos são configurados pelos seguintes pontos:

a. atribuição explícita, implícita, título de crédito (autoral) ou outros documentos indicando autoria, circunstâncias da criação – especialmente diários, correspondências, notas de editores ou registros legais;

b. evidências biográficas como: imputação de autoria, localização do indivíduo na época da publicação, ligações pessoais e afiliações políticas e/ou religiosas; e

c. história de atribuição de autoria anterior e as circunstâncias nas quais elas ocorreram.

Os índices intrínsecos se constituem dos pontos:

a. evidências estilísticas;

b. auto-referência ou auto-apresentação dentro do trabalho; e

c. evidências de temas, idéias, crenças e conceitos manifestados no texto.

O estabelecimento do grau de importância dos índices intrínsecos e extrínsecos variou com o decorrer do tempo e

com as aplicações de métodos de trabalho. Em 1960, por exemplo, Samuel Schoenbaum insistia que "*external evidence can and often does provide incontestable proof; internal evidence can only support hypothesis or corroborate external evidence*" (apud Love, 53). Hoje, essa certeza de Schoenbaum já não soa tão convincente. A maioria dos críticos aposta mais no exame dos índices internos de evidências observadas num texto, mas não procedendo do mesmo modo como se fazia no passado, quando essas análises eram impressionistas, e, sim, centrando atenção em trabalhos estilométricos, baseados em amostragens que envolvem, às vezes, milhões de palavras.

Em textos atuais, o estudo de autoria baseado em palavras raras utilizadas pelo autor ou, ao contrário, uso abundante de determinado vocábulo são características estilísticas, índices intrínsecos, que vêm se mostrando muito úteis para estabelecimento de autoria de um texto. Esse tem sido o caminho utilizado por Foster, por exemplo.

Outro ponto que se deve ter em mente ao começar a trabalhar nessa seara é a de que estudos dessa natureza são incompatíveis com o pensamento pós-estruturalista. Justifica-se esse axioma com a necessidade de se trabalhar sempre na crença de que a escrita traz em seu bojo características únicas para cada sujeito. E como aquelas teorias partem da proposta de construção do *eu*, este poderia ser criado à semelhança de outro, então a possibilidade de unicidade de estilo ficaria comprometida. Poderia haver clonagem do eu, do estilo, ou de ambos.

Dentro dessa visada de análise estilométrica, já tratamos de material textual constituídos por três casos diferentes: dois dos textos observados envolvem o ficcionista norte-americano, contemporâneo, Thomas Pynchon e o outro está constituído pela investigação da autoria de *Cartas chilenas.*

De Pynchon nos interessam os trabalhos de determinação da autoria das cartas de Wanda Tinasky, atribuídas ao ficcionista, e, relativamente à questão do estilo cíclico do autor, procuramos mostrar que a auto-reflexidade das obras está montada sobre uma lógica informática na distribuição das palavras *if*, *then* e *go to*. O terceiro caso é formado pelo debate em torno da, de todos conhecida, indefinição da autoria de *Cartas chilenas*. Esses casos de estudo indicados,

sobre o qual estabelecemos hipóteses de trabalho ainda estão, como projeto de trabalho, indefinidos, porque esperam financiamento para seu pleno desenvolvimento, mas temos já alguns dados preliminares para partilhar aqui sobre o material constituído pelas *Cartas chilenas*.

PARTE II – INÍCIO DE ESTUDO DE CASO: CARTAS CHILENAS

1. Introdução

Embora a idéia de fazer essa análise para determinação da autoria de *Cartas Chilenas* seja já de 2003, ainda não encontramos as condições favoráveis para desenvolvê-la. Não temos, ainda, como foi indicado, financiamento para a concretização desse projeto, mas nos anima a controvérsia que o tema poderá causar quando for divulgada através de eventos da mesma natureza deste em que estamos agora. Lançamos, por isso e nessa oportinidade, mais informações sobre as primeiras bases metodológicas que guiarão o trabalho e a apresentação de uma poucos resultados palpáveis já encontrados. Esses dados que enunciamos a seguir são algumas poucas amostras de importância, provavelmente, temporárias, resgatadas das primeiras simulações feitas com o objeto de estudo proposto.

2. Do Objetivo e Objeto

O objetivo primeiro da pesquisa é o de verificar a possibilidade de se obter índices quantitativos e/ou qualitativos na escrita que possam servir para evidenciar a autoria de textos anônimos ou apócrifos. Ou seja, determinar se a pessoa, ao escrever, deixa algum tipo de índice que possa servir como sendo um modo de impressão digital que singularize sua escrita. O primeiro objeto escolhido, per si, demonstra que profundidade e seriedade desejam-se impor a este projeto, tratando das *Cartas chilenas*. Elas são um conjunto formado por quatorze poesias satíricas que datam do século XVIII, atribuídas a Tomás Antônio Gonzaga, mas foram assinadas com um pseudônimo: o de *Critilo*. Existe muito para ser desvendado desse mistério formulado em torno da autoria desse texto: primeiro, as *Cartas* foram encontradas na forma de dois manuscritos diferentes, um contendo as sete primeiras cartas e outro as sete complementares; segundo, existem quatro manuscritos, ligeiramente diferentes, da

mesma época, contendo as sete primeiras cartas; e, como terceira questão, a *Epístola à Critilo* é atribuída a Cláudio Manuel da Costa.

De uma revisão bibliográfica desse tema resulta que vamos encontrar toda a crítica e história da literatura do século XX apontando Gonzaga como autor das treze cartas, da primeira à décima-terceira; de Cláudio Manuel da Costa seria a *Epistola*. Já a revisão do século XIX mostra vozes discordantes, como a de Silvio Romero (1980, p.429) que após uma rápida, mas convincente, análise de símiles encontradas nas cartas, indica Alvarenga Peixoto como provável autor das *Cartas*. O dito popular sobre opinião unânime é o que nos impulsiona a ir fundo nesta investigação e dela tirar lições que, se não resolverem a questão, pelo menos permitam caminhar no sentido de criar um mecanismo para averiguação tanto de textos anônimos como daqueles 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 passadas 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, utilizamos, preferencialmente, um software desenvolvido na Université de la Sorbonne Nouvelle – Paris 3, pela equipe CLA2T (**Cédric Lamalle**, **William Martinez**, **André Salem** e **Serge Fleury**), 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 nem sempre é tão amigável.

O LEXICO3 tem se mostrado uma ferramenta muito poderosa e de simples utilização. Ela nos permite, de forma

ágil, balizar livremente o texto a ser analisado, determinando como dividir o texto, fazer contagem das vezes que uma determinada palavra ocorre 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 à 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 disponível, muitos passos têm que ser, literalmente, adivinhados pelo pesquisador noviço durante a execução do trabalho.

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. Esta pequena porção 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, tamanho das palavras, riqueza do léxico, freqüência das palavras, uso da pontuação, freqüência de determinados sinais de pontuação, dentre outros. 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 (apud Peng e Hengartner, 2001) 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 de textos usando como material a ser privilegiado na verificação o número de palavras por sentenças utilizadas pelo autor. Outros pesquisadores usaram esses mesmos parâmetros em outros corpora, mas os resultados não parecem animadores.

5. Os Primeiros Parâmetros

Nossa primeira tentativa de aplicação de uma metodologia de trabalho específica para esse trabalho está sendo definida em termos da contagem e distribuição de palavras funcionais, tais como conjunções, dêiticos e preposições, existentes nas frases. A lógica que sustenta a escolha desse caminho é a de que essas palavras são usadas de forma mais independente e menos racionalizadas pelos autores na elaboração de um texto. As palavras funcionais são utilizadas, em determinado texto, de acordo com a necessidade de coerência textual e escolhida, geralmente, dentre um número reduzido de possibilidades. Essa escolha metodológica está justificada pelos resultados muito favoráveis obtidos por Peng e Hengartner (2001) na análise procedida sobre textos de Austen, Carther, Doyle, Dickens, Kipling, London, Marlowe, Milton e Shakespeare. Os pesquisadores indicados trabalharam com textos de autoria conhecida para determinar a eficácia do método de trabalho.

Como dissemos, as cartas estão divididas em dois manuscritos: um contendo as sete 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, o processo de trabalho aplicado foi o de cotejar os achados nas diversas *Cartas* com aqueles da décima-quarta. Após esse apanhado inicial, o propósito é o de comparar os achados em *Cartas Chilenas* com a produção de origem segura, isto é, a partir de textos cuja autoria está estabelecida, dos poetas contemporâneos ao do autor das *Cartas*. Neste primeiro momento estaremos cotejando

as contagens encontradas para as cartas segunda, décima e a *Epístola*.

6. Experimento 1

No exemplo citado, acima, desenvolvido por Peng e Hengartner (2001), foram utilizadas sessenta e nove 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 resultados relevantes foram:

Palavra	*segunda carta*	*décima carta*	*Epístola*	Palavra	*segunda carta*	*décima carta*	*Epístola*
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				

Observou-se, nesse primeiro levantamento, que *o*, *não*, *os*, *um*, *do*, *já*, *com*, *seu* e *da* apresentam um padrão que sugere estilos parecidos entre as *Cartas* segunda e décima, e divergentes em relação à *Epistola*. Já os léxicos *a*, *se* e *aos* remetem para outro tipo de interpretação.

A surpresa maior nessas primeiras simulações apareceu quando balizamos o LEXICO3 para que ele fizesse a contagem de segmentos repetidos em cada grupo de dez formas diferentes. O intrigante começou com o número de pares encontrados em cada uma das *Cartas*: na segunda, trinta e nove pares, na décima, quarenta e dois pares e na *Epístola*, vinte e três pares. Ou seja, o estilo do autor da *Epístola* é muito econômico quando se trata de segmentos repetidos, quase metade do encontrado nas outras duas.

Este *Experimento 1* revelou ainda uma informação interessante, levantamos que quatro pares de palavras ocorrem com relativa alta freqüência na tanto segunda e como

na décima *Cartas*, e não se fazem presentes, sequer uma única vez na *Epístola* . Essas palavras são:

Segmentos	*Segunda carta*	*décima carta*	*Epístola*
em que	3	4	0
o que	2	5	0
O nosso	5	4	0
que não	3	6	0

Os achados do *Experimento 1* que pareciam promissores nesse pequeno experimento ficaram diluídos, quase desapareceram, quando alargamos o horizonte de contagem para um maior número de *Cartas*, de forma que resolvemos abandonar esse viés para o momento. Mesmo porque, com esse aumento de corpora uma outra revelação nos chamou a atenção. Partimos para o *Experimento 2*.

7. Experimento 2

Começamos essa parte da pesquisa alargando os dados utilizados no *Experimento 1*, mas tentando a mesma metodologia usada por Peng para demonstrar que a palavras funcionais servem para identificar autoria de textos poéticos. O pesquisador, no entanto, não indica como chegou às palavras funcionais que ele utiliza em sua pesquisa. Neste segundo experimento, optamos por utilizar as 20 palavras funcionais mais freqüentes, retiradas de todos os blocos de poesia que seriam utilizados, mas que aparecesse em quantidade ≥ 7. Chegamos a essas palavras:

TABELA DE PALAVRAS FUNCIONAIS MAIS COMUNS*					
A	DE	ENTRE	NOSSO	QUE	UMA
AO	DO	LHE	O	SE	
AS	DOS	MAIS	OS	SEU	
COM	E	ME	PARA	SOBRE	
DA	EM	NA	POR	UM	

* As palavras não funcionais CHEFE, DOROTEU e MARÍLIA foram desprezadas devido as suas participações serem desequilibradas.

Já no início deste *Experimento 2*, definimos alguns parâmetros para efetuar as contagens que passamos, agora, a enumerar:

a. Apesar de Peng sugerir que os blocos ideais para a realização do experimento sejam com 1700 palavras, adotamos blocos de 1335 palavras pelo fato de a *Epístola* ter esse tamanho;

b. Desprezamos as 8ª e 13ª *Cartas*, por serem incompletas;

c. Os blocos de 1335 palavras foram tirados das partes iniciais de cada obra, como disponibilizada pelo NUPILL;

d. A palavra funcional *não* foi descartada devia a forte tendência às negativas presentes em as *Cartas*;

e. Foram utilizados como fonte para detecção de um possível padrão estilístico, textos reconhecidos como de autoria de Cláudio Manuel da Costa, Tomás Antônio Gonzaga, Silva Alvarenga e Alvarenga Peixoto;

f. Utilizamos como fonte para os dados de Tomás Antônio Gonzaga trechos de *Marília de Dirceu*; para Cláudio Manuel da Costa, blocos retirados de *Vila Rica*; de Silva Alvarenga, partes de *O desertor*; de Alvarenga Peixoto, *Poesias*.

Feita essas delimitações e tratamentos dos textos para deixá-los prontos para a análise com o LEXICO3, partimos para as contagens das palavras e gráficos de distribuição delas no texto, como sugere Peng, para concluirmos que não havia estabilidade nas palavras funcionais suficientes para fazermos qualquer proposição responsável. Muito provavelmente, trabalhamos com um universo de blocos muito pequeno (*Vila Rica*, três; *Marília*, três; *O desertor*, três; *Poesias*, três; e, mais um bloco de cada *Carta*). É pouco, para análises com ferramentas telemáticas, mas grande o suficiente para ser improvável sem elas. No total estávamos tratando corpora que somavam mais de trinta e duas mil palavras.

Notamos, entretanto, que havia uma estranha coincidência na ordem, nas contagens e nas freqüências em que as palavras que estávamos trabalhando apareciam na janela criada pelo LEXICO3. Para Cláudio Manuel da Costa, nos três blocos que estudamos as catorze primeiras palavras eram do tipo funcional e a décima quinta era uma palavra de carga semântica, por exemplo, observe-se o bloco 1:

1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	10ª	11ª	12ª	13ª	14ª	15ª
O	QUE	A	E	DE	SE	OS	DO	AO	MAIS	DA	AS	EM	UM	HERÓI

Para Tomás Antônio Gonzaga, em *Marília*, a coincidência ocorreu também, sendo que para ele a contagem acontecia na

décima primeira palavra, ou seja, após dez palavras funcionais a próxima era não funcional. Com Alvarenga Peixoto a contagem não foi tão exata, o primeiro bloco foi completamente discrepante diante das outras duas contagens: no primeiro bloco a palavra não funcional aconteceu na décima sétima posição, mas os dois blocos seguintes aconteceram na 10ª e 11ª posições, respectivamente. Na observação de Silva Alvarenga, verificamos que as palavras não funcionais ocorrerem nas 16ª , 16ª e 17ª posições, respectivamente. As *Cartas* se comportaram como no quadro abaixo

1ª	2ª	3ª	4ª	5ª	6ª	8ª	9ª	10ª	11ª	12ª	Epístola
13ª	13ª	18ª	12ª	13ª	15ª	16ª	21ª	17ª	14ª	19ª	12ª

Existem mais alguns dados que foram colhidos nessas contagens que podem colaborar para alguma conclusão. Chama a atenção, por exemplo, o número de formas (palavras diferentes) que aparecem em cada bloco de 1335 palavras. Ou seja, essa é uma medida que pode indicar o grau de extensão do vocabulário de cada um dos poetas. O poeta de *Cartas* apresenta grandezas diferentes para cada carta

1ª	2ª	3ª	4ª	5ª	6ª	8ª	9ª	10ª	11ª	12ª	Epístola
700	654	667	656	666	666	629	642	633	658	646	699

Esse dado não é conclusivo para definir o autor de *Cartas*, mas podem-se descartar alguns dos candidatos. Alvarenga Peixoto tem contagem superior a 700 formas nos três blocos utilizados nesta pesquisa (709, 727 e 712). Silva Alvarenga apresenta margem ainda mais dilatada: 745, 745 e 737. Como não acreditamos que alguém controle, conscientemente, a grandeza do vocabulário quando está escrevendo, apontamos para esse índice intrínseco como afirmativo para o descarte dos poetas citados neste parágrafo.

Cláudio, em *Vila Rica*, apresenta seguinte contagem de formas para os três blocos que trabalhamos: 719, 678 e 689. Enquanto que Tomás, em *Marília*, apresenta a seguinte contagem: 602, 606 e 664.

Antes de partirmos para uma conclusão dessa experiência, vale ainda destacar um outro indício que devemos deixar registrado. Na observação da utilização de sinais gráficos, salta à vista a utilização de pontos de

interrogação. Em *Vila Rica*, Cláudio utiliza, no total, apenas 8 sinais de interrogação e todos no segundo bloco. Já em *Marília*, Tomás faz uso do sinal da seguinte forma: bloco 1, 7; bloco 2, 28; e, bloco 3, 3. Já em *Cartas* temos o resultado distribuído assim:

1ª	2ª	3ª	4ª	5ª	6ª	8ª	9ª	10ª	11ª	12ª	Epístola
8	7	9	9	1	1	12	18	17	1	6	3

Um outro dado que nos apareceu nesse *Experimento 2* e que merece registro trata-se da palavra mais utilizada por cada um dos poetas nos blocos analisados e nas *Cartas*. Essa observação para Silva Alvarenga, mostra que os léxicos *e, a* e *o* são os mais freqüentes, enquanto o léxico *que* aparece na quarta posição. Quando observamos os blocos de Alvarenga Peixoto, nota-se que o *que* aparece em segundo lugar em dois blocos e em terceiro no terceiro bloco. No tocante a Cláudio, em *Vila Rica*, o *que* aparece como segundo mais freqüente em dois blocos (precedido de *o*, uma vez, e *a*, em outro), no terceiro bloco o *que* aparece em quarta posição (precedido de *a, o* e *e*). A observação dos três blocos de *Marília*, revelou o *que* aparecendo na segunda posição em dois blocos e em primeiro no último bloco. Em *Cartas*, das treze analisadas, em doze o *que* aparece em primeiro lugar de freqüência. A única exceção é a *Carta* de número seis. Nela o vocábulo ocorre no segundo lugar.

Nas simulações que realizamos, conseguimos identificar algum padrão numérico quando analisamos os versos de poesias de autoria conhecida. Isso é verdade quando observamos a ordem em que aparecem as palavras não funcionais, também o é quando tratamos da dimensão do vocabulário utilizado pelos poetas, o uso de determinados sinais gráficos, ou, ainda, as palavras mais freqüentes utilizadas pelos poetas. Mas quando voltamos os olhos para *Cartas Chilenas*, não sentimos confiança para indicar qualquer autoria.

8. Conclusão

Continuaremos com nossa investigação em busca de uma maneira de definir o real autor de *Cartas Chilenas*, se possível. Mas, para esse momento, optamos por indicar que o que acontece com essas poesias é exatamente o que Love (2002) indica como sendo um dos maiores problemas par definição de autoria: múltiplos autores.

Preferimos, neste momento, acreditar na imagem que nos vem à mente toda vez que lemos *Cartas*. Ela é a de um grupo pequeno, de três ou quatro jovens poetas, em uma taverna de iluminação rarefeita, feita com candelabros e velas, bebendo vinho e se divertindo enquanto compunha versos satíricos que ridicularizavam o governante local. Essa é nossa crença atual.

Referências

ALVARENGA, Manuel Inácio da Silva. **O desertor**. In: Universidade Federal de Santa Catarina; NUPILL. Disponível em: <**www.cce.ufsc.br/~nupill**>. Acesso em 14 de outubro de 2004.

ÁVILA, Affonso. **O lúdico e as projeções do mundo barroco**. 2. ed. São Paulo: Perspectiva, 1980.

COSTA, Cláudio M. **Vila Rica**. In: Universidade Federal de Santa Catarina; NUPILL. Disponível em: <**www.cce.ufsc.br/~nupill**>. Acesso em 14 de agosto de 2004.

FERREIRA, Delson. **Cartas chilenas**: retrato de uma época. 2. ed. Belo Horizonte: Ed. UFMG, 1986.

FOSTER, Donald. **Author Unknown**: Tales of literary detection. New York: Henry Holt, 2000.

GONZAGA, Tomás. **Cartas chilenas**. In: Universidade Federal de Santa Catarina; NUPILL. Disponível em: <**www.cce.ufsc.br/~nupill**>. Acesso em 10 de abril de 2003.

GONZAGA, Tomás. **Marília de Dirceu**. In: Universidade Federal de Santa Catarina; NUPILL. Disponível em: <**www.cce.ufsc.br/~nupill**>. Acesso em 10 de abril de 2003.

LOVE, Harold. **Attributing authorship**: An introduction. Cambridge: Cambridge U. P., 2002.

O'DONNELL, Bernard. **An analysis of prose study to determine authorship**. Paris: Mouton, 1970.

OLIVEIRA, Tarquínio. **As cartas chilenas**: fontes textuais. São Paulo: Editora Referência, 1972.

PEIXOTO, Inácio José de Alvarenga. **Poesias**. In: Universidade Federal de Santa Catarina; NUPILL. Disponível em: <**www.cce.ufsc.br/~nupill**>. Acesso em 10 outubro de 2004.

PENG, R. e Nicolas HENGARTNER. *Quantitative analysis of literary styles*. **The American Statician**, v. 56, n. 3, p. 175-185, 2001.

ROMERO, Silvio. **História da literatura brasileira**. 7ª ed., Rio de Janeiro: José Olympio, 1980, tomo 2. 5 tomos.

SELLIN, E. e FOHRER, G. **Introdução ao Antigo Testamento**. v. 2. São Paulo: Paulinas, 1977.

[1] Devo essa informação, uma leitura crítica muito útil e várias conversas sobre o assunto à Profa. Dra Ana Maria Koch. Aproveito este espaço agradecer por sua disponibilidade e amizade.