Em Busca do DNA de um texto

José Augusto Da Silva Neto[1]

[1]Departamento de Informática e Estatística – Universidade Federal Do Piauí (UFPI)
CEP 64.049 - 550 – Teresina – PI - Brasil

j_augusto_n@hotmail.com

Abstract. *The study's search for authorship of a text gains an important ally, the software Lexico 3. This powerful tool helps researchers from different fields and dealing with texts, showing a range of characteristics of the corpus analyzed. The present study aimed to examine thoroughly the software in question, concomitantly quest for authorship of work Cartas Chilenas, from lexicometric and stylometry studies, given the data collected. And thus it was possible to indicate a methodology for such research, and demonstrate the relevance of using a computational tool for this process.*

Resumo. *O estudo da busca de autoria de um determinado texto ganha um importante aliado, o software Léxico 3. Esta poderosa ferramenta auxilia pesquisadores de diferentes domínios e que lidam com textos, evidenciando uma gama de características dos corpus analisados. Diante disso, o presente trabalho objetivou perquirir o software em questão, concomitantemente à busca de autoria da obra Cartas Chilenas, a partir de estudos lexicométricos e estilométricos, frente aos dados coletados. E dessa forma, foi possível apontar uma metodologia para este tipo de pesquisa, bem como comprovar a relevância do uso de uma ferramenta computacional para tal processo.*

1. Introdução

Com o desenvolvimento de ferramentas telemáticas para tratamento de dados em grande volume, como a de mineração, por exemplo, pôde-se criar outras possibilidades de trabalho, específicas e novas, agora também direcionadas para a área de determinação de autoria de textos. Através do uso raro de determinadas palavras pelo autor ou, ao contrário, uso abundante de determinado vocábulo, definem-se algumas características estilísticas, isto é, índices intrínsecos que se mostram úteis para o desenvolvimento desse estudo.

Diante desse contexto, o projeto "Em busca do DNA de um texto" visa fomentar, através do uso de ferramentas telemáticas, o estudo que envolve a determinação da autoria de textos apócrifos. Para isso, foi utilizado, preferencialmente, um *software* desenvolvido na *Université de la Sorbonne Nouvelle* – Paris 3, pela equipe CLA2T, denominado LEXICO3. E a partir da aplicabilidade que esse programa proporciona foi possível traçar os seguintes objetivos: compreender o funcionamento do *software* escolhido, possibilitando a elaboração de um manual em língua portuguesa, com as principais funções que o programa disponibiliza; digitalizar e balizar - adequar os textos à formatação exigida pelo programa - todos os textos utilizados nas análises, bem como outros *corpus*, que servissem de base de dados para projetos futuros; verificar a possibilidade de se obter índices quantitativos e/ou qualitativos na escrita que servissem como evidências a autoria de textos anônimos ou apócrifos; aferir a eficácia

de ferramentas telemáticas na determinação de autoria desses textos e apontar uma metodologia para definição de autoria a partir de estudos lexicométricos e estilométricos; evidenciar a autoria da obra *Cartas chilenas*, conjunto formado por quatorze poesias satíricas que datam do século XVIII, atribuídas a Tomás Antônio Gonzaga, mas que foram assinadas com um pseudônimo: o de Critilo.

2. Metodologia

O LEXICO3 é um programa de aplicação lexicométrica extremamente versátil e de utilização não muito complexa. Ele tem se mostrado uma ferramenta muito poderosa que nos permitiu, de forma ágil: o balizamento do texto a ser analisado, determinando como dividi-lo; fazer contagem das vezes que uma determinada palavra ocorre dentro de um *corpus*; determinar o tamanho de um segmento repetido a ser pesquisado; fazer o levantamento das ocorrências dos segmentos repetidos; indicar a distribuição das palavras dentro do texto; expor as concordâncias que ocorreram com uma palavra; elaborar gráficos indicando as freqüências relativas e absolutas da aparição de uma palavra em uma determinada baliza, ou ainda, utilizando cálculos específicos demonstrar uma "inflação" da partícula analisada em um determinado texto, em comparação ao restante do *corpus* analisado.

Entretanto, diante de uma avaliação heurística, baseado nos critérios de usabilidade de Nielsen, verificou-se a fragilidade do *software* quanto à prevenção de erros, bem como quanto à disponibilização de mecanismos de ajuda, haja vista a existência de um manual que não se mostra totalmente compreensível e ainda não abrange todas as ferramentas do programa. Diante disso, a frente inicial de trabalho concentrou-se no entendimento preciso das funções que o programa dispõe, através da leitura robusta dos manuais prévios do *software* em questão, escritos em língua inglesa e francesa. Concomitantemente, foi possível a elaboração de um manual em língua portuguesa, com as principais instruções de uso desse *software*, desde a sua instalação até a utilização das ferramentas de análises lexicométricas que ele disponibiliza.

Após esse minucioso trabalho, fora dado início ao processo de digitalização de alguns textos utilizados para testes iniciais. Foram eles: três cartas, escolhidas aleatoriamente, da autora Wanda Tinasky, além de dois textos do autor Thomas Pynchon. Em seguida, esses textos passaram por um processo de balizamento, de acordo com as exigências, quanto ao formato de arquivo (.txt), bem como outras regras de formatação textual impostas pelo *software*. Dessa forma, esses textos utilizados como controle para análise lexicométrica, permitiram as primeiras experiências quanto à analogia entre autores distintos, como também o aperfeiçoamento na utilização das regras do balizamento e maior praticidade no uso das ferramentas do Lexico3, evitando assim erros futuros.

De posse das informações necessárias para o manuseio do *software*, passou-se a trabalhar com o texto proposto para esse projeto, a obra Cartas Chilenas. Além desta, digitalizou-se e balizou-se alguns conjuntos de poemas de escritores que, por serem contemporâneos a Tomás Antônio Gonzaga, poderiam ser autores das Cartas. Foram eles: Silva Alvarenga, Cláudio Manuel e Alvarenga Peixoto. Em seguida, todos os textos foram reduzidos a blocos de 1335 palavras, com exceção da Epístola, 8ª e 13ª cartas, que possuem *corpus* inferior a esse tamanho.

Trabalhou-se em duas frentes: análise quantitativa e qualitativa. Em um primeiro momento, na análise quantitativa, determinaram-se quais eram os elementos, ou formas, que seriam investigados. Algumas possibilidades que se apresentaram foram as seguintes: freqüência das palavras, distribuição destas no *corpus*, tamanho das frases, riqueza do léxico, uso da pontuação e freqüência de determinados sinais de pontuação.

Diante disso, o primeiro passo foi a contagem e distribuição das cem palavras mais freqüentes de cada bloco analisado, através das seguintes ferramentas do léxico 3: PCLC (*Principales Caractéristiques Lexicométriques*), SP (*Spécificités*). Em posse dos dados fornecidos pelo Lexico 3, utilizou-se o programa Microsoft Office Excel 2003 na elaboração de gráficos mais didáticos, no formato de barras.

Fazendo-se uso da mesma seqüência de passos, diferenciando-se apenas pelas informações exigidas, foi possível a feitura de diversos gráficos. Estes abrangeram, além das cem palavras mais freqüentes de cada bloco analisado, as seguintes informações: palavras funcionais (concentrou-se em artigos, preposições, conjunções) mais utilizadas; quantidades e tipos de verbos encontrados; comparação entre os textos analisados, quanto à utilização de certas palavras funcionais; análise quanto à riqueza de vocabulário de cada bloco, através do dado *Formes* (Quantidades de formas gráficas diferentes utilizadas em cada texto) etc.

Por meio dessas informações, foi possível comparar os textos entre si, e assim evidenciar proximidades lexicométricas entre as Cartas Chilenas e os possíveis autores, quanto ao uso ou desuso de certas palavras ou expressões. Para tanto, utilizou-se a ferramenta *Statistiques par partier*, que gera gráficos contendo as freqüências relativas e absolutas daquelas. Os gráficos abaixo demonstram aquela analogia, através da análise do dado *formes*, bem como das freqüências relativas das palavras "que", "a" e "o":

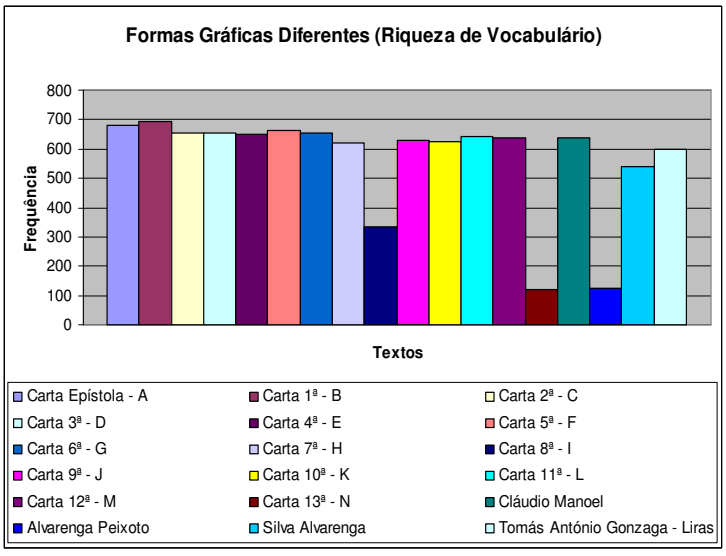

Figura 1: Gráfico que mostra a riqueza de vocabulário presente em cada texto analisado.

Figura 2: Gráfico das freqüências relativas de algumas palavras funcionais presentes em cada bloco analisado.

3. Resultados

Com o conhecimento adquirido acerca da utilização do Lexico3, foi possível a elaboração de um manual prévio, em português, com o detalhamento instrucional das suas principais ferramentas. Além disso, comprovou-se a importância do uso desse *software*, como relevante contribuinte na determinação da autoria de textos apócrifos, permitindo a caracterização de uma metodologia para definição de autoria a partir de estudos lexicométricos e estilométricos, mediante uma ferramenta computacional.

4. Conclusões

Ressalta-se que todo o trabalho de busca da autoria das Cartas Chilenas está sendo facilitado pelo uso do *software* em questão. E diante dos resultados coletados das análises feitas até o momento, verifica-se como esse texto apresenta características similares a um número significativo de autores.

Referências

Ferreira, D. (1986). Cartas chilenas: retrato de uma época. Belo Horizonte: Ed. UFMG, 2ª ed.

Gonzaga, T. (2006). Cartas chilenas. Companhia de bolso. São Paulo, Schwarcz Ltda.

Peng, R. e Hengartner, N. (2001). Quantitative analysis of literary styles. The American Statician. V.56, N 3, p. 175-185.

Dictionary The American Heritage Dictionary for Windows. 3. Ed. Version 3.6a, 1994.

Brandão, S. (2005). Atribuição de Autoria: um problema antigo, novas ferramentas. Texto Digital, n°2.

Amstel, F. V. (2003) As 10 heurísticas de Nielsen. http://www.interacts.com.br/10_heurísticas_de_nielsen.shtml.

Nielsen, J. (1993). Usability Engineering. New York, Academic Press.