FRONTEIRAZ
ISSN 1983-4373

Revista do Programa de Estudos Pós-Graduados em Literatura e Crítica Literária da PUC-SP

nº 18 - julho de 2017

http://dx.doi.org/10.23925/1983-4373.2017i18p98-111

Aspectos introdutórios para uma crítica numérica da literatura

Introductory aspects for a numerical critique of literature

Saulo Cunha de Serpa Brandão *

RESUMO
Neste ensaio, pretendo explorar as possibilidades investigativas da crítica numérica (SADIN, 2015). Começamos transitando por diversos tipos de *software*, dando maior ênfase àqueles de distribuição livre e disponíveis à comunidade, mas comentamos também alguns outros produtos comercializados. Em seguida, passamos a comentar as diversas possibilidades críticas à disposição do pesquisador a partir dos dados numéricos obtidos e que, com um pouco de conhecimento do que representam os valores estatísticos, podem revelar muito dos estilos de autores e de correntes literárias. Outra vertente que aflora com os números obtidos dos textos literários é a da identificação de autoria de textos apócrifos ou escritos sob pseudônimo. Apresento também dois *softwares* de minha lavra; são eles: O NEOLO e o *Xfragment*. Ambos são escritos em Python; o primeiro tem como funcionalidade principal a identificação e extração de neologismos de textos em português ou em inglês, além de 16 outras funcionalidades; o segundo é pontual e faz a identificação e filtragem de fragmentos no texto. Por enquanto, *Xfragment* está capacitado para trabalhar somente com textos em língua inglesa.
PALAVRAS-CHAVE: Crítica Numérica; Estilometria; NEOLO; *Xfragment*

ABSTRACT
In this essay, I intend to explore the investigative possibilities of numerical criticism (SADIN, 2015). To begin with, I go through several kinds of software, placing more emphasis on those that are freely distributed and available to the whole community; I will also comment about a few other products found in the market. Next, I will discuss the various critical possibilities available to researchers, based on the numerical data obtained, which, with a little knowledge of what statistical values represent, can reveal

* Universidade Federal do Piauí – UFPI – Programa de Pós-graduação em Letras – Teresina – PI – Brasil. saulo@ufpi.edu.br

FRONTEIRAZ
ISSN 1983-4373

Revista do Programa de Estudos Pós-Graduados em Literatura e Crítica Literária da PUC-SP

nº 18 - julho de 2017

a great of the style of authors and literary trends. Another critical aspect that surfaces with the data obtained from literary texts is the identification of authorship of apocryphal texts or texts written under a pseudonym. I will also present two kinds of software that I have developed myself, namely: NEOLO and Xfragment, both of which written in Python. The main functionality of the former is the capacity to identify and extract neologisms from texts in both Portuguese and English, but which also has 16 complementary functionalities; the latter is more specific, as it identifies and filters fragments in texts. So far, Xfragment has been prepared to operate only with texts in English.
KEYWORDS: Numeric Criticism; Stylometry; NEOLO; Xfragment

A tentativa de aproximar o estudo da literatura aos preceitos adotados pelas ciências, ditas duras, vem de muito tempo. Já na virada do século XIX para o XX, teóricos russos defendiam a necessidade de afastar os estudos literários do subjetivismo que reinou durante boa parte do século XIX. Influenciados pela fenomenologia husserinana, esses pensadores lançaram-se à busca do que pensavam ser o fenômeno na arte literária. Eles empreenderam estudos valiosos sobre a estrutura de contos maravilhosos, o ritmo das poesias, os aspectos relevantes da sonoridade do léxico. Hoje, teóricos como Terry Eagleton (2005) apontam os erros cometidos pelos formalistas russos na compreensão do que, de fato, era o fenômeno na literatura.

A tentativa era (e é, ainda) de mostrar a necessidade que a comunidade de teóricos e críticos literários tem de se conceber um *modus operandi* que privilegie um olhar objetivo e com procedimentos e axiomas claros de como analisar o texto literário. O parágrafo anterior é um exemplo desse movimento, mas podemos também trazer à baila as escolas que sucederam os russos, como a Escola de Praga, Grupo Tel et Quel, estruturalistas e semioticistas[1].

Na atualidade, os estudiosos da Crítica Numérica fazem suas pesquisas buscando os mesmos objetivos, mas com outro modo da ciência como padrão e fugindo da impossibilidade da fenomenologia. A busca deles é aplicar um procedimento caro às ciências duras, mas com comportamento um pouco simplista de considerar o texto literário como o fenômeno, evitando discussões filosóficas a respeito dessa abordagem, e buscar a descrição do fenômeno com dados matemáticos e estatísticos. Alguns desses dados são de fácil lavra para obtenção, mas outros exigem cálculos muito complexos que dependem de manipulação por profissionais estatísticos e matemáticos. Isso deixa com o crítico numérico a competência para analisar os resultados. Mas, mesmo para essa análise, o crítico tem que entender a operação estatístico-matemática que foi empregada.

Para executar a análise numérica, utiliza-se *softwares* especializados em fazer a mineração dos textos em questão e oferecer como resultados dados numéricos que devem ser avaliados pelo pesquisador. Obviamente, para que os resultados façam sentido, o estudioso tem que estar familiarizado com as ferramentas disponibilizadas pelo programa em uso. Nada adianta simplesmente inserir o corpus em análise no

[1] Essas escolas são citadas, mas não esquecemos de falar do caso alemão, que definiu, desde o começo do século XX, que a disciplina amplamente conhecida como Teoria da Literatura teria como título Ciência da Literatura. Mas é uma história um pouco mais longa e este é o motivo de eu me contentar com as proposições mais difundidas por aqui.

software. O resultado disponibilizado não significará muito para um analista despreparado.

O que acontece na prática é que os especialistas em literatura fazem uso muito singelo de programas de análise de texto. Isso acontece por falta de familiaridade do literato com programas de computador e com as tecnologias em geral; desconhecimento do mínimo necessário de matemática e, desconhecimento de estatística. Por essas fraquezas, quase sempre, quando mostramos a outros especialistas em literatura nossa análise, estes olham com indiferença para o trabalho. São poucos os *outsiders* que reconhecem na prática da crítica numérica um viés rico para investigação de textos literários.

Outro aspecto da análise computacional de textos literários é que quase nunca encontramos um *software* que responda a todas as necessidades de um projeto de pesquisa. Portanto, é fundamental que o pesquisador trabalhe com um cardápio de *softwares* e saiba o que cada um desses programas pode oferecer para a pesquisa desejada.

A pesquisa numérica apresenta, além das dificuldades citadas acima, problemas muito sérios quando falamos do tratamento do texto. Os melhores *softwares* trabalham com o modo *.txt.* Mas, antes da formatação do modo, temos que fazer limpezas dos textos quase *ad infinitum*. Limpar, nesse contexto, significa descobrir os erros de grafia que ocorrem quando da transformação do texto, que geralmente nos chega em PDF, para *.txt*. Por melhor que esteja o texto, ele sempre vai apresentar um sem número de erros e alguns difíceis de se encontrar. Cito um erro muito comum: a palavra em PDF é *como*, mas após a transformação ela se torna *corno*. A dificuldade vem pelo fato de a palavra *corno* existir em português e os corretores automáticos não indicá-la como um erro. Esse tipo de problema impede que o pesquisador faça correções automáticas. Ele pode até fazer uma correção automática em parte do texto, mas sempre sobra muito material para ser limpo manualmente. Alguns teóricos e até pesquisadores experientes calculam que o trabalho de limpeza do texto pode tomar até 60% do tempo destinado ao projeto de pesquisa.

Quanto à plataforma usada para rodar os programas, podemos distinguir dois tipos de programas. Os que estão hospedados em um *site*, caso em que, para rodá-lo, o pesquisador precisa fazer um *upload* do *corpus* e, às vezes, marcar opções de análises desejadas. Esses programas têm algumas vantagens, como: estar em plataforma *Windows*, que é conhecida pela maioria das pessoas, e não ocupar espaço no disco do

usuário, mas apresenta algumas desvantagens: rodam *corpus* de tamanho limitado e apresentam problemas (*bugs*) com frequência. Os defeitos ocorrem porque os programas, na maioria, são escritos por não especialistas e a interface com *Windows* ou *HTML* é de difícil construção. Alia-se aos problemas já mencionados o fato de a maioria não passar por revisões técnicas periódicas.

Um exemplo de um mecanismo nos moldes do descrito e criticado acima pode ser testado na plataforma Contador de Palavras, criado pelo Grupo de Lingüística (sic) Insite. O *site*[2] é interessante para fazer demonstrações rápidas e para instigar alunos a se interessarem pelo tipo de crítica tratado neste ensaio. Nesse programa, o usuário apenas faz um *upload* do texto que deseja analisar e dá um *click* no botão *Processar* e imediatamente terá o resultado processado. Como indicado acima nas desvantagens, podemos aprender na página do programa que ele só processará textos de até 12 *kbytes*. Mas traz um número muito grande de resultados que, se bem interpretados, podem fornecer dados suficientes para fazer a análise de um texto com muita proficiência. Esse programa, diferente da maioria dessa natureza, está completamente funcional. Mas, como explicado, se o crítico não souber o que está buscando, os números apresentados representam muito pouco. Uma saída para se poder utilizar esses programas *on-line* com textos um pouco maiores é transformar os textos-objeto para o formato com radical *.txt* (textos desestruturados obtidos com programas como o *Notepad, Notepad++, Jedit, Sublime* ou outros similares).

Dentre os programas que são baixados de páginas da internet, destaco o *Lexico3*[3], de origem francesa. Esse programa é bastante intuitivo e fornece resultados muito precisos. Ele revela dados simples como dicionário de palavras, lista de palavras pela frequência, concordância das palavras, segmentos repetidos, faz gráficos diversos, analisa o texto completo ou partições do texto. Nessa ferramenta, pode-se comparar diferenças estatísticas entre partes do texto. E outra vantagem é que roda textos muitos grandes e em línguas diferentes. Para que o programa aceite idiomas diferentes, precisa-se tratar o texto para a língua desejada. O processo é simples, o texto tem que ser salvo em codificação própria da língua (*auto-detect* – UTF-8 – UTF-16 – USASCII etc.). O *Lexico3* é de distribuição gratuita e foi desenvolvido na *Université de la Sorbonne Nouvelle* – Paris 3.

[2] Disponível em: [http://linguistica.insite.com.br/corpus.php].
[3] Disponível em: [http://lexi-co.com/].

Outro *software* que chama a atenção pela complexidade e também gratuito é o *Semantic Search Engine*[4]. Ele foi atualizado até 2014 e trata-se de uma ferramenta poderosa que tem como principal função agrupar elementos da narrativa em torno de TROPES. Ele faz a análise do texto e revela em uma estrutura gráfica a conexão existente entre as mais destacadas construções textuais. Dessa forma, podemos verificar a proximidade relacional entre personagens, entre lugares, palavras frequentes, fazendo as conexões possíveis dos elementos da narrativa. O programa, apesar da última atualização ter sido em 2014, é funcional. Mas não é uma ferramenta de fácil utilização e também é de difícil interpretação, porque os resultados são fornecidos em uma página muito carregada de informações. Trata-se de um programa profissional e demanda tempo para o manuseio e interpretação.

Outro *software* do segundo tipo que merece menção aqui é o *WordNet*[5]. Ele é muito utilizado no mundo acadêmico e seu funcionamento aparenta o de um dicionário em meio digital, sendo que as informações geradas são muito mais sofisticadas do que um dicionário. Os autores o definem como um "[...] *large lexical database of English*".[6] As palavras das diversas classes se relacionam dentro das combinações possíveis e geram sinônimos para cada contexto possível. O *WordNet* faz a desambiguação das palavras lexicalmente idênticas, ou quase idênticas, e, a partir da proximidade das palavras de relações possíveis, é capaz de dar o conceito de cada uma dessas palavras baseado no relacionamento inter-palavras que elas aceitam. Ele foi desenvolvido por pesquisadores da Linguística de Corpus e aqueles do Processamento de Língua Natural.

Sobre os programas comercializados, tratarei do *Hyperbase* e do *Alceste*. Os dois programas tiveram origem comum[7], mas seguiram, depois, por caminhos diferentes e têm funcionalidades muito parecidas. O primeiro ficou limitado ao mundo acadêmico, sendo comercializado por preços muito módicos (minha licença custou cerca de €$300,00), mas atualmente ele é distribuído livremente[8]. O programa tem diversas e complexas ferramentas, das quais destaco um gráfico bidimensional, mas que simula uma figura tridimensional que relaciona as estruturas do texto literário, mostrando proximidade ou afastamento dos elementos. O gráfico é gerado a partir de dados estatísticos e matemáticos que são relacionados e exibidos em forma arbórea. Além

[4] Disponível em: [http://www.semantic-knowledge.com/].
[5] Disponível em: [https://wordnet.princeton.edu/].
[6] Fonte: FELLBAUM, C.; TENGI, R. *What is WordNet?* Disponível em: [https://wordnet.princeton.edu].
[7] SYLED-CLA2T (*Système Linguistiques Énonciation Discursivité - Centre d'Analyse Automatique des Textes*) na *Université de la Sorbonne Nouvelle* – Paris 3.
[8] Disponível em: [http://ancilla.unice.fr/].

dessa funcionalidade complexa, o programa ainda pode revelar: riqueza lexical, evolução do vocabulário, a distância ou conexão do texto e a coloração do tema. Uma desvantagem nesse *software* é que ele trabalha com bases de literatura fixas, existindo a base francesa, a base latina, a base portuguesa (graças ao trabalho do Prof. Carlos Maciel) e a base algeriana. Fornece também dicionários de referência específicos para francês, inglês, italiano e português. O programa aceita *corpora* na casa dos milhões de palavras.

O Alceste[9] tem funcionalidade parecida com o *Hyperbase*, mas, por ter sido pensado para o mercado de gerenciamento de informações de *corpora* gigantes e por grandes organizações da área da comunicação, marketing e para análise sociológica e política, ele tem a sua base aberta, podendo analisar textos diversos. Minha licença do *software* custou o equivalente a US$2.800,00. Embora com funcionalidades parecidas com as do *Hyperbase*, ele tem uma interface mais amigável e profissionalmente desenhada. Hoje, a Alceste tem escritórios espalhados pelo mundo e uma clientela majoritariamente corporativa.

Trabalhando com esse tipo de *software*, o pesquisador tem uma carta de possibilidades investigativas muito ampla e diversificada. Comentarei três trabalhos desenvolvidos no Núcleo de Pesquisa em Literatura Digitalizada (NUPLID) da Universidade Federal do Piauí (UFPI). As pesquisas foram feitas para o desenvolvimento de dissertações apresentadas na pós-graduação em Letras da instituição.

Os agora mestres que desenvolveram a pesquisa sob minha orientação são: Diego Meireles, Samara Liz e Lívia Guimarães. O principal *software* utilizado foi o *Lexico3*, mas também foram utilizadas outras ferramentas, inclusive o *Excel* da família do *Office* da *Microsoft*.

O primeiro desenvolveu um trabalho para desmistificar crenças sobre as características da poesia de H. Dobal, poeta piauiense de destaque no cenário nacional, que eram tidas como certas pelo senso comum no Piauí. A *intelligentsia* local tinha, há muito, concordado e difundido a ideia que o léxico usado por Dobal era pobre (número limitado de formas). Diego desenvolveu um trabalho exemplar, mostrando que o vocabulário usado pelo poeta é simples, mas não pobre. A riqueza lexical das poesias de

[9] Disponível em: [http://www.alcestesoftware.com.br/].

Dobal equivale à de outros poetas seus contemporâneos, como Manuel Bandeira. Diego fez a mesma comparação com diversos outros poetas, o que ratificou a descoberta.

Samara Liz, por ser uma estudiosa da Literatura Portuguesa, desenvolveu um trabalho muito instigante: ela fez a comparação estilométrica entre a poesia de Fernando Pessoa com a de alguns heterônimos. A descoberta foi interessante porque revelou que o ortônimo e os heterônimos não diferem apenas no nome e nos interesses (o futurista, o pastor, o místico), mas em seu estilo, riqueza lexical e outros dados reunidos a partir de núcleos de estro.

Já o estudo desenvolvido por Lívia foi no sentido de mostrar que, subjacente à poesia pessimista de Augusto do Anjos, havia uma outra poesia que tinha uma visão mais favorável da vida. A pesquisa foi desenvolvida buscando os estados emocionais que envolviam a palavra amor na poesia de Augusto. A pesquisa, de novo, desautorizou o que o senso comum diz sobre a dicção do poeta. Inicialmente, fez-se um trabalho procurando por amor, suas desinências e seus *stem* (variações, desinências com o mesmo radical) e também sinônimos fortes do termo amor. O resultado confirmou duas hipóteses com as quais trabalhávamos: primeiro, o autor trabalhava proativamente para ser considerado um poeta maldito; e, segundo, ele queria se inscrever dentro da filosofia pessimista schopenhaueriana.

O procedimento foi, grosso modo, analisar o livro *Eu*, que é composto de poemas escolhidos e publicados pelo próprio Augusto em vida. Após sua morte, um amigo próximo amealhou as poesias inéditas e voltou a publicar o livro com o título *Eu e outras poesias* e, posteriormente, *Todas poesias*. Então, quando analisamos os dados colhidos do livro *Eu*, de fato, o registro é claro em relação ao pessimismo do autor, mas quando analisamos os resultados gerados a partir do livro *Eu e outras poesias*, o perfil do autor muda e uma face apaixonada, mais bem-humorada aparece. E essas características ganham ênfase quando rodamos as poesias publicadas postumamente isoladamente.

Existem mais alguns projetos de pesquisa de natureza textométrica sendo desenvolvidos no NUPLID. Alguns alunos de graduação estão desenvolvendo uma pesquisa promissora com peças teatrais reconhecidamente de Shakespeare e comparando com outras em que há dúvida sobre a autoria e também com textos de outros autores contemporâneos ao bardo inglês. Eles descobriram que Shakespeare tem uma assinatura muito estável na questão das letras mais frequentes utilizadas. Isso se confirma em todos os textos dele que analisamos, sejam comédias, tragédias altas e

baixas e históricos. Mas esse estudo ainda demanda melhores investigações e o planejamento de um protocolo de pesquisa para referendar os resultados que obtivermos mais adiante.

Temos mais duas pesquisas em desenvolvimento por pesquisadores mais experientes de nosso grupo, inclusive eu. Estamos analisando o léxico utilizado por Guimarães Rosa para construir *Grande sertão: veredas*; e o léxico utilizado por Thomas Pynchon na escrita de *Mason & Dixon*. O projeto busca avaliar o que de fato é neologismo nas obras, pode-se perceber diferença nos estilos dos discursos masculinos e femininos nos romances e, também, comparar o perfil dos discursos dos gêneros nas ficções com os apreciados nas línguas ordinárias portuguesa de variação brasileira e inglesa de variação norte-americana.

Um outro viés de pesquisa muito explorado por quem trabalha com esses programas é a verificação de autoria. Segundo autores como Love (2002), Foster (2000), Kenny (1992) e até o bem mais recente Franco Moretti (2007), cada autor tem uma marca textual registrada. Como eu já mencionei acima, Shakespeare tem uma marca indelével quanto às letras usadas com maior frequência. Outro exemplo que encontrei em uma pesquisa que fiz em 2014-2015, em fase de finalização do relato dos resultados, foi sobre a predileção de Thomas Pynchon pela palavra *even*. Ela é, muitas vezes, mais frequente na ficção pynchoniana do que seu uso no inglês ordinário e mais frequente também do que nos textos não-ficcionais do autor. Então, a parte mais difícil desse tipo de pesquisa é estabelecer qual é a marca do autor. Feito esse achado, podemos estabelecer os parâmetros da pesquisa.

Desde 2003, eu venho labutando para esclarecer a autoria das *Cartas chilenas*. Não é tarefa fácil, porque estamos muito distantes temporalmente da escrita delas e no Brasil não existe(ia) a cultura de preservar documentos e tampouco a de historiar o caminho seguido pelo texto. Essas carências terminam por impor muita dificuldade para o pesquisador. Em um primeiro experimento que montei em 2005, conclui que o autor da *Epistola* é de fato Cláudio Manuel da Costa. Mas, no conjunto, ou carta por carta, não consegui vinculá-las a Tomás Antônio Gonzaga, que figura como o autor no cânone dos estudos literários no Brasil, nem a nenhum outro poeta árcade brasileiro (Cláudio Manuel da Costa, Silva Alvarenga, Alvarenga Peixoto, Basílio da Gama, Frei Santa Rita Durão etc.). Os meus achados não corroboram essa assertiva canônica, que já foi ratificada por pessoas do peso de Manuel Bandeira, Cecília Meirelles e Afonso Arinos.

Na atualidade, além do trabalho que faço com alguns colegas sobre *Grande sertão* e *Mason & Dixon*, desenvolvo uma pesquisa solo em que tento desvincular o nome do ficcionista norte-americano Thomas Pynchon do de Vanda Tinasky. Na década de 1980, apareceram cartas publicadas em um jornal californiano com o mesmo tom sarcástico/humorístico de contracultura característico de Pynchon. Logo, alguns estudiosos da literatura pynchoniana comentaram que existia a possiblidade de o escritor ser o autor das cartas, que vinham assinadas como Vanda Tinasky – *The lady bag*, mas nenhum especialista abraçou a hipótese. No meio jornalístico apareceram vários atores que tentaram provar a tal vinculação. Por fim, o Prof. Donald Foster (2000), já referenciado, encerrou a disputa, negando indubitavelmente a possível autoria de Pynchon para *The letters of Wanda Tinasky* (1996). Esse achado de Foster – ele narra a história no livro acima citado – aconteceu por coincidência e baseado em fatos históricos externos, sem nenhuma utilização de análise textual. O meu propósito é provar que os textos não são da produção de Pynchon a partir de achados internos ao texto. A pesquisa está muito avançada e devo publicar um artigo na revista *Orbit* no primeiro semestre de 2018.

Voltemos, agora, o olhar sobre a crítica numérica com o propósito de determinar autoria e seu estado no mundo. E aqui trataremos de outros imbróglios relativos a essa crítica.

Não se pode dizer que a crítica numérica é vanguarda nesses anos do começo do século XXI. Sabemos que desde o medievo monges católicos se debruçavam sobre manuscritos de autoria duvidosa[10] para, por meio de coincidências numéricas, tentar revelar seus autores, inclusive inserindo ou não partes na *Bíblia*. Tampouco as experiências com a crítica numérica só aconteceram no medievo; são famosas as pesquisas feitas acerca da obra shakespeariana. O leitor interessado pode encontrar denso levantamento histórico sobre o estado da arte da crítica numérica autoral consultando o livro de Anthony Kenny, *The computation of style* (1982), ou, ainda mais completo, o de Harold Love, *Attributing Authorship: An introduction* (2002).

Em se tratando de problemas autorais envolvendo o bardo inglês, encontramos especial relevo no trabalho desenvolvido por Donald Foster (2000), que, em sua tese doutoral sobre a autoria do poema choroso *A funeral elegy to master William Peter*, atribuiu a produção a William Shakespeare. Sua tese seria sobre a autoria de sonetos

[10] Existem diversos evangelhos apócrifos, como os de Tomé, Judas, Maria Madalena.

atribuídos ao bardo, mas, com a descoberta de uma elegia inédita, que poderia ser de Shakespeare, o pesquisador se voltou para esse problema.

A grande crítica feita pelos *experts* em Shakespeare à época das hipóteses defendidas por Foster é que não existe suporte exterior ao texto que apoie a proposta. Ou seja, nada se sabia se Shakespeare era conhecido das pessoas envolvidas com o poema, tampouco ele esteve nos locais mencionados na história da elegia. A poesia era de baixíssima qualidade em sua dicção, diferente da produção de Shakespeare. A proposta de Foster estava baseada nos índices internos ao texto: signos linguísticos, aliterações, sequência de palavras repetidas nos textos sabidos de Shakespeare e que apareciam também na elegia. Enfim, este exemplo aconteceu na década de 1980, mas não está completamente esquecido. Aqui e ali ainda aparecem comentários ou atualizações sobre o debate.

O método matemático-estatístico defendido por Foster se apresentou muito efetivo na descoberta do autor do romance *Primary colors*, publicado anonimamente, mas o pesquisador rapidamente produziu um artigo indicando o jornalista Joe Klein como autor. Este negou a autoria por algum tempo, até que caíram nas mãos de outros jornalistas partes de escrita cursiva do romance. Estes compararam os manuscritos com os escritos jornalísticos de Klein e concordaram com o que Foster havia definido algum tempo antes. Depois disso, Klein reconheceu que a obra era de sua autoria.

O pesquisador da Vassar University, Foster, voltou aos holofotes da mídia quando ajudou o FBI a definir a identidade do terrorista doméstico norte-americano, conhecido como Unabomber, como sendo o matemático Theodor John Kaczynski, depois que este publicou um manifesto de 35 mil palavras. Ou seja, material suficiente para que um pesquisador especialista em lexicometria e estilometria[11] identificasse o autor do *corpus*.

No livro de Foster existem outros quatro casos em que ele conseguiu identificar autores para textos apócrifos, anônimos ou escritos sob pseudônimos ou heteronímia.

O fato é que muito tempo depois de defender sua tese, Donald Foster teve que se render a achados que davam completo suporte para a não vinculação da elegia à Shakespeare. O pesquisador não tentou replicar, apenas veio a público e concordou que, com as novas descobertas, ele não podia mais dar suporte a sua própria tese.

[11] Termos usados para denominar o mundo de quem usa estratégias matemáticas-estatísticas para analisar textos. Em nosso caso, textos literários. Existe também um termo mais generalista para quem se envolve com a crítica numérica, que é 'textometria'.

O leitor que quiser saber mais sobre outras histórias de textos que tiveram seus autores definidos por meio da lexicometria e estilometria deve ler o livro de Love já mencionado. Ele conta, inclusive, um evento do século XVII em que a única prova da inocência de um marinheiro era um bilhete. Por fatos como esses, hoje existe uma especialização em Linguística Forense.

Passemos à última parte do texto para falar brevemente sobre dois programas que criei nos últimos dois anos e meio. Primeiro, darei alguns detalhes sobre o NEOLO. Este programa foi escrito em *Python*. Ele é de código aberto e gratuito. E, atualmente, ele roda em plataformas *Linux*. Tem como função principal retirar de textos as palavras que sejam neologismos (vêm também palavras estrangeiras e estrangeirismos). O funcionamento dessa ferramenta é simples de descrever: o usuário insere na sintaxe de comando de NEOLO o nome do texto que quer analisar e, em alguns segundos, recebe um arquivo em *.txt* com os resultados. O texto escolhido vai ser comparado a dicionários e listas de palavras com mais de 200 mil verbetes. Como o programa, os dicionários que devem ser anexados a ele são completamente abertos. Podem ocorrer eventuais erros em que palavras conhecidas são selecionadas como neologismo e, nesse caso, o usuário pode inseri-las no dicionário em uso e elas não mais aparecerão na listagem de resultado. Esse tipo de erro tem sido bastante comum por conta da última reforma da Língua Portuguesa, mas estamos trabalhando muito para resolver esse problema.

NEOLO também faz seis verificações de riqueza vocabular. Elas são descritas em um artigo por Torruella e Capsada (2013). Três delas são sensíveis ao tamanho do texto, ou seja, para que o resultado de uma análise de riqueza seja aceito com essas três ferramentas, os textos comparados devem estar com o mesmo número de palavras. As outras três ferramentas são de base estatística e podem comparar a riqueza textual de textos de tamanhos diferentes. A capacidade de estabelecer com a maior precisão possível a riqueza vocabular de um texto é muito importante para verificar a sua autoria.

As outras ferramentas disponíveis no NEOLO[12] são as seguintes: contagem de sentenças no texto, tamanho médio das sentenças, lista de neologismos, o tamanho do texto em formas[13], tamanho do texto em tipos, número de *hapax legomenon*,

[12] O programa pode ser obtido em: [https://github.com/jcrowgey/neolo].

[13] Formas são todas as palavras que aparecem em texto e Tipos são as que se repetem para formar o número de formas. Mais adiante aparecerá o termo *hapax legomenon* que significa as palavras que aparecem apenas uma vez no texto. Esses termos são importantes para o cálculo da riqueza lexical.

distribuição de palavras pelo seu número de letras, lista de tipos com sua frequência, lista de *hapax legomenon* e média do uso de pontuação.

O segundo *software* é o *Xfragment*. Esse programa também foi escrito em Python e tem como função extrair e contar o número de fragmentos existentes em um texto. Fragmento é, aqui, conceituado como uma sequência de palavras com sintaxe correta, mas sem verbo. Para o *Xfragment*, uma sentença começa com letra maiúscula e termina com as seguintes pontuações gráficas: ponto, exclamação, interrogação e reticências. O programa passa por um processo de etiquetagem para reconhecer as palavras e suas classes gramaticais. É um processo de treinamento que realizamos várias vezes até perceber-se que já não existem erros grosseiros. No processo, também existe uma adaptação para as possibilidades sintáticas com as classes gramaticais. O *Xfragment* foi pensado para avaliar numericamente a assertiva comum dos teóricos de que os textos pós-modernistas são fragmentados. O programa está ainda em fase de teste, mas já foi utilizado em uma pesquisa para comparar os textos de John Barth e Thomas Pynchon e apresentou resultados muito bons. Por enquanto, o programa está sendo utilizado exclusivamente para textos em língua inglesa. Mas está em nossa agenda para o segundo semestre de 2017 começar a fazer a aclimatação dele para o português. Em breve poderemos incorporar o *Xfragment* no dia a dia de nossas pesquisas.

Concluo esse ensaio com a certeza de que os estudos numéricos vão ser usados cada vez mais. Isso vai acontecer nos próximos poucos anos porque existem já alguns pesquisadores no Brasil interessados nessa perspectiva de pesquisa e, consequentemente, estão introduzindo esses programas nos projetos de alunos do PIBIC e alunos de pós-graduação. Estes, por sua vez, continuarão utilizando *softwares* em pesquisas mais complexas porque é muito gratificante fazer pesquisas em que os resultados são palpáveis, longe de elucubrações interpretativas abstratas produzidas por pesquisadores sérios e compromissados eticamente com o mundo da pesquisa ou de outros que produzem textos macarrônicos para fingir erudição, mas completamente descompromissados com qualquer nível de honestidade intelectual.

REFERÊNCIAS

EAGLETON, T. *Literary theory:* an introduction. 10. ed. Oxford: Blackwell Publishing, 2005.

FACTOR, T. R. (Ed.). *The letters of Wanda Tinasky.* San Francisco: Vers Libre Press, 1996.

FELLBAUM, C.; TENGI, R. *What is WordNet?* Disponível em: [https://wordnet.princeton.edu]. Acesso em 19 fev. 2016.

FOSTER, D. *Author unknown:* on the trail of anonymous. New York: Henry and Holt Ed., 2000.

KENNY, A. *The computation of style.* New York: Pergamon Press, 1982.

LOVE, H. *Attributing authorship:* an introduction. Cambridge: Cambridge U. P., 2002.

MORETTI, F. *Graphs, maps, three:* abstracts models for literary history. New York: Verso Editors, 2007.

TORRUELLA, J.; R. CAPSADA. Lexical statistics and typological structures: a measure of lexical richness. *Procedia: Social and Behavioral Sciences*, 95, 2013.

Data de submissão: 23/03/2017
Data de aprovação: 04/05/2017