TEXTO DIGITAL
Revista de Literatura, Linguística e Artes

A necessidade faz a oportunidade: software NEOLO

Need makes the opportunity: software NEOLO

Saulo Cunha de Serpa Brandão[a]

a Universidade Federal do Piauí, Teresina, Brasil - brandaosaulo@yahoo.com

Palavras-chave:
NEOLO. Textometria. Neologismo. riqueza lexical

Resumo: Neste trabalho disserto sobre uma pesquisa que venho desenvolvendo desde 2013 e que resultou na construção de um software para analisar diversos dados de textos. O programa conta com 12 funcionalidades, mas neste ensaio eu tratarei apenas das 4 mais complexas e inéditas em algoritmos para uso por pesquisadores sem experiência em computação.

Keywords:
NEOLO. Textometry. Neologismo. Lexical Richness.

Abstract: In this work, I will present a research that I have been developing since 2013, which resulted in the construction of a software to analyze various texts data. The program has 12 features, but, in this essay, I will deal only with the 4 most complex and unprecedented in algorithms for the use by researchers with no experience in computing.

1 INTRODUÇÃO

Início este ensaio informando que trataremos da história de uma pesquisa, que resultou na construção de um software que denominei NEOLO. A investigação está inserida no viés teórico-crítico da textometria, em suas duas possibilidades de visadas: lexicometria e estilometria. Neste viés, o pesquisador trabalha com o uso de computadores e programas próprios para levantamento de variáveis estatísticas do texto e também com planilhas para registro e cálculos.

Antecipo, por didaticamente necessário, que o software NEOLO conta com 12 (doze) ferramentas funcionais, mas pela economia textual necessária, tratarei neste ensaio das 4 (quatro) ferramentas mais complexas que foram implantadas no programa.

No ano de 2013, eu saí de minha instituição – Universidade Federal do Piauí - para cumprir um estágio sênior na University of Washington – Seattle, o estágio foi planejado para durar 1 (um) ano, mas que no final me ocupou por um ano e meio. Para essa missão, eu contei com apoio financeiro da CAPES[1], pelo qual sou muito grato. Cheguei em Seattle no começo de agosto de 2013 e antes de 15 (quinze) dias eu já estava devidamente regularizado na universidade e também instalado, com apartamento, carro, seguros e tudo mais em ordem. De forma que na segunda quinzena de agosto, eu já estava na labuta, na pesquisa.

2 O PROBLEMA

O projeto que eu desenvolvia tinha o título: *Neobarroco e Pós-modernismo: investigando relações*. É obvio que se fosse tratar do assunto a partir de elementos sócio-históricos o problema já estaria resolvido, ou melhor, não haveria problema. Apenas a localização geográfica desses dois fenômenos já traria claras indicações dos porquês de seus aparecimentos. Enquanto o pós-modernismo apareceu na América do Norte em uma situação onde, por exemplo, o consumismo estava em seu apogeu (décadas de 1960 e 1970); existia uma tensão entre Estados Unidos e o Vietnam, que depois viraria uma das guerras mais nonsense em que os EUA se envolveram; a guerra fria também estava no seu ápice; a população dos EUA era muito jovem e irreverente, com isso não aceitavam as políticas

[1] Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Fundação CAPES

impostas pelo seu governo, exemplo disto é movimento *hippie*; a marcha do reverendo Martin Luther King Jr.; os direitos civis tinham sido igualados por lei, mas na prática guetos continuavam existindo, principalmente em relação aos negros, mas logo outras diferenças vieram à tona como o preconceito contra orientais, na costa oeste, preconceito contra os latinos, na fronteira imensa com o México etc.

O Neobarroco, por outro lado, aparece com o boom da literatura latino-americana na mesma época que o Pós-modernismo aparecia no Norte. O cenário em que essa literatura acontece não se coaduna com o descrito no parágrafo anterior. Muito pelo contrário, os países do sul eram conhecidos como parte do 3º mundo, ou seja, eram parte de um mundo que não tinha saído de um estágio muito rudimentar do capitalismo e este já enfrentava um movimento político situado mais à esquerda, de tendência socialista. Insatisfeitos com este movimento, as nações capitalistas, na economia, e de direita enquanto posicionamento político (que nas décadas seguintes se transformaria em uma posição neoliberal) agiram apoiando militares em diversos países para que esses tomassem o poder em seus respectivos países[2]. Então, era um continente praticamente dominado por ditadores de direita e manipulados pelo grande capital internacional. Como consequência vem o aparecimento de guerrilhas, urbanas ou não, tentando desestabilizar os regimes ditatoriais.

Dito isto, volto a questão de como tratar a pesquisa uma vez que pelo viés sócio-histórico a pesquisa seria nula, ou melhor: não haveria o quê pesquisar, como já dito. A proposta do projeto nasceu durante aulas que ministrava sobre literaturas Fantásticas, Realismo Maravilhoso e Realismo Mágico. Neste contexto, minhas leituras me levaram a relação entre os textos das 2 (duas) últimas correntes ao Neobarroco.

A partir desse ponto, eu me preocupei em estudar o aparecimento e as características do Neobarroco e as que me chamaram a atenção foram as que indicavam a utilização em abundância de vocábulos e a construções sintáxicas intrincadas. Estas herdadas ou copiadas

[2] Para uma visão panorâmica do que aconteceu no Brasil, por exemplo, recomendo o documentário *Um ano que durou 21* de Camilo Tavares, 2013. Sobre a América Latina, recomendo *As veias abertas da América Latina*, de Eduardo Galeano, e sobre a disposição de os EUA interferirem na política da A.L., visite o sítio da comissão Nacional da Verdade sobres a *Operação condor*, em <http://www.cnv.gov.br/index.php/2-uncategorised/417-operacao-condor-e-a-ditadura-no-brasil-analise-de-documentos-desclassificados>. Eu o visitei em 25/12/2015.

do barroquismo do Setecentos. Estas características eram tentadoras para quem estuda textometria (lexicometria e estilometria).

Sobre a proliferação de palavras e a sintaxe complexa Monsiváis (2009, p. 189) vai justificar como um *horor vacui, or a fear of empty spaces*, ou seja, de um horror ao espaço em branco, da necessidade de continuar narrando sem conseguir atingir a satisfação do texto completo. Voltarei para discutir melhor esse aspecto da pesquisa.

A pergunta principal da pesquisa apareceu porque eu tinha um bom conhecimento sobre o Pós-modernismo, uma vez que tinha tratado do assunto com vagar enquanto desenvolvi minha tese doutoral. E durante a pesquisa que tinha desenvolvido no final da década de 1970, encontrei alguns críticos que falavam também da hipertrofia, tanto em número como em complexidade, do léxico e da sintaxe.

Essa dupla informação sobre duas escolas literárias que, por razões geográficas, deveriam estar muito apartadas e serem completamente distintas, traziam características tão intimas, porque é o tricô do texto, que eram coincidentes.

Outro aspecto que merece cuidado é que as 2 (duas) escolas literárias apareceram no mesmo momento da história, sendo que uma no hemisfério Norte das Américas e outro no hemisfério Sul. Elas respondiam ou nasciam de cenários sócio-políticos muito diferente, mas ambas tinham a característica estilística de ter um léxico abundante e sintaxe complexa.

Depois que constatei essa semelhança e passei a revisitar alguns textos sobre o Pós-modernismos que faziam menção ao Neobarroco. Esse flashback me fez perceber que o problema da semelhança/diferença entre Neobarroco e Pós-modernismo estava quase intocado, quando muito, teóricos dedicavam 1(um) ou 2(dois) parágrafos sobre o assunto, sem muita elucubração para argumentação, logo, o assunto estava esperando alguém que chegasse a um denominador comum de o que pertence do Pós-modernismo e o que é do Neobarroco.

Vejamos agora algumas opiniões bem acreditadas no mundo dos teóricos da literatura sobre Pós-modernismo Vs. Neobarroco. Assim posto, vejamos algumas das posições:

a. Omar Calabrese (1992), por exemplo, afirma que o Pós-modernismo virou um rótulo que se coloca em obras de arte, mesmo que ela não atenda vis-à-vis as

características do Pós-modernismo. Ele sugere que o termo Neobarroco seria mais qualificado para abarcar todas as expressões artísticas da Pós-modernidade;

b. Já Barth (citado por Linda Hutcheon, 1992) e Severo Sarduy (1975) sugerem que deve o Neobarroco ser inscrito na cultura hispânica e não no Pós-modernismo;
c. Hutcheon (idem), discute Carlos Fuentes e Gabriel Garcia Marquez, que são considerados autores vinculados ao Neobarroco, junto com escritores outros do Pós-modenismo como Thomas R. Pynchon e John Barth sem fazer distinção;
d. Kelly A. Wacker (2007), respondendo à pergunta se o Neobarroco seria outro termo para denominar o Pós-modernismo? Ela responde dizendo que todo Neobarroco é pós-modernista, mas nem todo Pós-modernismo é Neobarroco (Esta é também a minha posição sobre o assunto, embora não a considere final);
e. Kelly A. Wacker (2007) fala de uma tendência que existe em considerar o Neobarroco como *Post-modern Baroque*.

No desenvolvimento da pesquisa encontrei, também, teóricos que buscaram explicar os nós sintáxicos na construção dos textos das duas correntes. Aí, encontrei mais uma coincidência: as propostas vão se apoiar no pensamento do teórico do psicanalista francês Jaques Lacan.

Jameson, em seu trabalho seminal *Postmodernism or, the cultural logic of late capitalism* (1999), quando escrevendo sobre o estilo no Pós-modernismo, compara-o com o modelo de comunicação da esquizofrenia, segundo o psicanalista francês J. Lacan. O teórico americano aponta que no modelo da patologia existe uma fratura na cadeia de significação. Ou seja: *When the links of the signifying chain snap, then we have schizophrenia in the form of a rubble of distinct and unrelated signifiers.* (p. 26). O teórico continua seu raciocínio para depois dar como exemplo a música de John Cage. Ele fala da desconfortável sensação que a audiência fica diante dos silêncios impostos nas partituras. Ao ponto de quando a nota seguinte aparece, o sujeito já mal se lembra da nota anterior e daí o mal-estar está estabelecido.

Ele continua sua argumentação para afirmar que essa fragmentação tende às teorias da diferença que

> ...stress disjunction to the point at which the material of the text, including its words or sentences, tend to fall apart into random and inert passivity, into a set of elements which entertain separations from one another. (p. 31)

Dentro desse panorama, o que sobra dessa falha de comunicação? Ruído? Com certeza sobra uma abundância de palavras e sentenças que não se coadunam, a sintaxe fica comprometida, a comunicação fica falha, quiçá anulada.

De outra parte, vamos observar sintomas semelhantes nos textos neobarrocos. Severo Sarduy (1975) reconhece a abundância de vocábulos, e teoriza sobre o fenômeno a partir, também, do psicanalista francês Jaques Lacan. Ele identifica a "verbiage" como sendo uma característica do Neobarroco, senão vejamos:

> An overflowing cornucopia, renowned for its prodigality and dissipation…a mockery all functionality of all sobriety... hence, its mechanism of periphrasis, of digression and detour, of duplication and even of tautology. Verbiage, squandered forms, language which, because of its excessive abundance, can no longer designate things but only other designators of things [signifiers] which enfold other [signifier] in a mechanism of signification which ends by designating only itself, revealing its own grammar. (p. 118)

Ambos os teóricos, Jameson e Sarduy, argumentam que existe uma tentativa de comunicação em um jogo de relações entre significantes, e não entre signos que termina por comprometer a comunicação. Isto deixa o leitor com a permanente sensação de frustração diante do texto.

3 DESENVOLVIMENTO

O projeto no estado que eu descrevi até agora foi submetido a uma das mais importantes pesquisadoras do Neobarroco no presente, a Profa. Monika Kaup, da University of Washington – Seattle. Ela concordou supervisionar o meu trabalho em uma missão de pós-doutorado. Logo após, submeti o projeto à CAPES e recebi o apoio da agência para essa missão que foi de Estágio Sênior. Desnecessário, mas a supervisão de Kaup só durou 2 (dois) meses. Ela achava que eu era um aluno de doutorado, muito autoritária. A supervisão passou para o visionário Dr. Walter Andrews, do departamento de Civilização do Oriente Médio. Ele coordena um grupo de pesquisa muito ativo em desenvolvimento de softwares para tradução da língua persa (do início do século XX) para o inglês e outros projetos menores.

Na continuidade, passei a desenvolver meus trabalhos utilizando diversos softwares, dentre eles: Lexico 3, Aintconc 3.2, Calibre, AintFileConversor, as diversas ferramentas do TAPoR, além de algumas ferramentas do Hyperbase e do Alceste[3].

Os resultados que obtive me deixaram razoavelmente satisfeito, cheguei à conclusão que os textos pós-modernistas e os neobarrocos, efetivamente, são muito mais ricos em número de léxicos do que os textos de outras correntes literárias. Eles apresentam um número exagerado de hápax legomena (palavras que aparecem apenas uma vez no texto). E o que posteriormente ficou claro para mim é que muitos desses hápax eram palavras inventadas pelos autores. A partir dessa descoberta passei a fazer uma seleção semiautomática dos hápax que não eram reconhecidos pelos dicionários. Conclui duas coisas no processo: 1º - eu estava fazendo um trabalho insano e sem qualquer certeza de que trataria um número de neologismos que fosse representante da população, então minha pesquisa ficaria sem o devido caráter científico; e, 2º - minha proposta era a de trabalhar com utilização de softwares e da maneira que eu estava conduzindo dependia muito de meu trabalho manual; isso descaracterizava a pesquisa.

4 A SOLUÇÃO: NEOLO

Passei alguns dias debruçado sobre o problema e decidi que a saída para o dilema era construir um software que pudesse fazer o grande trabalho automaticamente. Passei a pensar e definir como seria esse programa. Novamente encarei um problema que minha ignorância sobre programação. Eu sabia o que queria, mas não sabia se era possível proceder as operações algoritmicamente. Nesse momento eu procurei um programador profissional da UW que tinha virado um bom colega e extremamente disponível para minhas dúvidas. O nome dele é Stacy Waters e além do conhecimento sobre desenvolvimento de software ele é também doutor em literatura inglesa. E dessas conversas e minhas necessidades nasceu o *blueprint* do programa chamado NEOLO.

O *script* e as estatísticas foram escritos por Joshua Crowgey. Ele era aluno de doutorado em Linguística de Corpus e diretor de tecnologia do grupo de pesquisa coordenado por Walter Andrews. A ideia inicial era extrair os neologismos de textos em inglês e português. Mas no

[3] Esses dois últimos foram menos utilizados por dois motivos: são muito difíceis de manipular (não existem manuais) e eles não são intuitivos (friendly use), além disso, na época, os dois eram softwares comerciais. Hoje, o Hyperbase é de distribuição gratuita.

desenvolvimento do trabalho muitas outras funcionalidades foram incorporadas ao programa, inclusive 3 (três) ferramentas para medir a riqueza lexical em textos de tamanhos diferentes. Dentre estas, existe uma ferramenta que tem o acrônimo HD-D. Esta ferramenta é extremamente complexa e, durante a feitura, dependemos de muitas horas para entender os procedimentos estatísticos e várias outras muitas horas para implementar a ferramenta (o mérito é exclusivamente de Joshua. Eu servi de incentivador!). No total, Neolo conta, hoje, com 12 ferramentas, algumas simples, mas úteis, e outras mais complexas.

Importante dizer neste momento que NEOLO foi escrito em Python e que roda exclusivamente em Linux. Mas pretendemos num futuro próximo colocar o programa para rodar on-line. E, em outro momento, mais no futuro, tê-lo uma versão para Windows.

Para ter acesso ao NEOLO, o primeiro passo é ter uma máquina de Linux instalado em uma partição do disco rígido. Neste espaço, o usuário deve instalar o Python (se já estiver instalado computador). Feitas essas tarefas iniciais, se o pesquisador for usar o Linux pela primeira vez e ainda não faz uso do GitHub[4], ele deve proceder da seguinte maneira:

1 - Na linha de comando do Linux, digitar: $> sudo apt-get install gig, depois:

2 - $> git clone https://github.com/JCrowgey/neolo para rodar NEOLO deve-se usar a seguinte chave:

3 - ./neolo + os gatilhos de NEOLO que interessam à pesquisa em curso[5]

Na sua funcionalidade principal, a lógica do programa depende da criação de filtros na forma de bancos de palavras. Assim, tivemos de fazer alguns bancos de palavras em português e em Inglês. Os bancos em cada língua têm em seu conteúdo: dicionário, lista de palavras, nomes próprios, abreviações mais comuns. Estes bancos ficam alojados dentro de pastas separadas para cada um deles e elas são colocadas dentro de uma pasta principal em que está salvo também o algoritmo do programa NEOLO. Para cada tipo de pesquisa o pesquisador pode invocar qualquer uma das pastas, ou todas/parte elas. Ou, ainda, podem ser criados outros bancos de palavras de interesse da investigação.

[4] O GitHub.com é uma plataforma para desenvolvimento colaborativo de softwares.
[5] No final, está inserido um anexo com as diversas possibilidades de exame por parte de NEOLO.

Em termos de dicionário para o português utilizamos 2 (dois): o 1º *Dicionário da Língua Brasileira* (1832), e o 2º *Novo Dicionário da Língua Portuguesa*[6] (1913), e para o inglês o disponível foi o *Webster's Unabridged Diccionary* (2009). Em termos de banco de palavras gerais, o NEOLO conta, ainda, nas duas línguas, de listas de palavras que foram criadas para suprir as companhias de telecomunicação com material linguístico que os celulares e outros *gadgets* usam para sugerir palavras para o usuário e corrigir palavras digitadas com erros de grafia. Há também listas para nomes próprios e das abreviaturas mais comuns nas línguas (Mr., Sr.). Esses corpora têm, em cada língua, um total aproximado de 400.000 (quatrocentos mil) verbetes.

Para uma pesquisa que estamos tocando neste momento e com publicação dos resultados a serem prevista para o 1º semestre de 2018, criamos um banco de palavras com centenas de vocábulos dicionarizados da variação da língua portuguesa falada nos interiores de várias regiões do Brasil e mais 1 (um) corpus contendo fauna e flora. Essa pesquisa nasceu pela necessidade de elucidar o que é neologismo ou não em livros como Grande sertão: Veredas de João Guimarães Rosa. O resultado surpreende. Além dos dicionários e listas citadas acima, tivemos de criar dois corpora de 1.000.000 (um milhão) de palavras aproximadamente, cada 1 (um), que serve de corpus testemunha para cada língua. Estes corpora foram formados obedecendo certas características defendidas pelos especialistas (e.g. Aluísio e Almeida, 2006) em Linguística de Corpus. Esses corpora-testemunha servem para validar alguns resultados e/ou cotejar as estatísticas do texto ficcional com a língua, dita, ordinária.

As outras duas listas, nomes próprios e abreviaturas, que criamos, têm funções primárias semelhantes e as funções secundárias diferentes. Estas pastas são utilizadas quando o pesquisador quer encontrar detalhes sobre pontuação e/ou número de palavras por sentença. Isto porque que sentença para o NEOLO é um conjunto de formas[7] (palavras) que inicia com uma letra maiúscula e termina em: interrogação, exclamação, ponto ou reticências. Se rodarmos o programa sem esses filtros, NEOLO interpretará que os nomes próprios serão inícios de sentenças e que os pontos das abreviaturas seriam pontos finais. Estes erros trariam

[6] Esses dicionários foram mantidos inalterados em seus arquivos originais e gerei cópias que passaram por um procedimento para atualizar a grafia das palavras. Então quando eles são invocados por NEOLO, se apresentam nas duas versões. O dicionário em inglês não sofreu alteração por ter sido publicado em data recente.

[7] O vocabulário especializado da Linguística de Corpus adota a palavra "formas" para se referir ao todo das palavras que aparecem no texto, inclusive suas repetições. Por outro lado, chamam de "tipos" a ocorrência única de cada palavra (como se fosse a lista de verbetes do texto).

resultados falsos quando a avaliação fosse sobre as estatísticas de sentença e pontuação (sintaxe). Em sua função secundária, esses bancos de palavras são usados para retirar os nomes próprios e abreviaturas se eles forem interferir negativamente nos resultados da pesquisa.

Resta ainda dizer que no 2º semestre de 2017, com a ajuda do pesquisador Adiel Mittmann (UFSC), nós pudemos inserir nos processos de trabalho de NEOLO, dois gatilhos novos que são *stemmings*, ou seja, com eles o soft passou a reconhecer palavras que tenham radicais comuns conhecidos. Eis um exemplo banal, mas didático: a palavra **tábua** está no banco de palavras de NEOLO, mas **tabuleiro** não está. Neste caso, **tabuleiro** seria listado como neologismo, sem ser. Com a introdução do *stemming*, NEOLO passa a identificar as duas palavras. A utilização desse recurso trouxe uma economia de tempo substancial para o pesquisador. O mesmo Adiel está empenhado em disponibilizar o nosso programa para rodar na internet.

NEOLO é um programa muito maleável porque nos permite adaptá-lo para as necessidades da pesquisa, retirando gatilhos ou os inserindo na chave[8] que define como ele deve proceder. Além disso, as listas de palavras, nomes próprios e abreviaturas estão em constante crescimento. Toda vez que ele, NEOLO, lista uma palavra dicionarizada no meio dos neologismos, o pesquisador deve acrescer o novo vocábulo no corpus que lhe é pertinente e submeter de novo o texto à análise do programa.

5 FERRAMENTAS DE DETERMINAÇÃO DA RIQUEZA LEXICAL

Descrita essa função primeira do Software NEOLO, passaremos a tratar de outro aspecto das análises Lexicométricas e Estilométricas que são difíceis para o pesquisador da área de Letras pela pouca, ou nenhuma, afinidade dele com as coisas da informática e estatística. Trata-se do cálculo da diversidade lexical, ou riqueza vocabular (ou lexical), na pesquisa comparativa[9].

[8] Colocarei um anexo com uma tabela dos gatilhos possíveis, presentemente, e outras informações.

[9] Quase sempre o pesquisador de Letras terá de fazer comparações entre 2 (dois) textos quando se trata de riqueza lexical, porque os números achados em um texto não significam nada se não comparados com os de outro texto. Muitas vezes a comparação é com corpora testemunha, ou seja: corpora que tentam representar a língua falada ordinariamente.

A dificuldade ocorre devido aos corpora trabalhados serem de tamanhos diferentes (em número de formas). Porque a característica geral dos textos é que não medida que eles se alongam, os indicadores de riqueza diminuam. Isto porque o cálculo da riqueza dá-se, via de regra, dos cálculos comparativos entre número de tipos Vs. número de formas. Uma vez que na medida que número de formas aumenta, diminui assimetricamente o número de tipos.

Trocando em miúdos: quanto menor o texto, mais rico em léxico ele é. Atente para a fórmula mais simples de calcular a riqueza: TTR= t/n[10] (Templin, 1957). A riqueza neste caso é a divisão do número de tipos pelo número de formas. Como o número de formas não pode jamais ser menor do que o número de tipos, o resultado sempre é menor ou igual (≤) a 1 (um). Observe o seguinte exemplo:

O rato roeu a roupa do rei Artur. Neste caso nós temos 8 (oito) tipos e 8 (oito) formas, então o índice de riqueza nesse caso é 1 (8/8 = 1), ou seja, é a maior possibilidade de riqueza lexical.

Já em: **O rato roeu a roupa do rei Artur. A roupa que o rato roeu era de gala.**

Formas = 17

Tipos = 9 então: 9/17 = 0,52

O segundo exemplo apresenta riqueza lexical de 0,53 (cinquenta e três centésimos). Praticamente metade da riqueza lexical do 1º exemplo.

Usar essa fórmula sem a devida observância do tamanho dos textos vai acarretar em resultados errados e inservíveis.

A tendência dos pesquisadores da área de Letras que não têm formação em estatística é trabalhar limitando o tamanho dos textos para poder fazer a comparação da riqueza lexical de corpora de tamanhos diferentes. Assim, se esse pretenso pesquisador tiver um corpus A com 15.000 (quinze mil) formas e o outro corpus B contar com 12.000 formas, ele resolverá o problema retirando 3.000 (três mil) formas do corpus A.

[10] TTR = Type-Token Ratio, ou seja, Relação Tipo-Forma. Refira-se à nota 6, para os conceitos.

Acontece que quando ele faz a supressão de formas no corpus A, ele vai introduzir no sistema um índice subjetividade indesejada. No momento que ele escolhe retirar os vocábulos das primeiras páginas do corpus, ou da última parte do corpus, ou, ainda, do meio do corpus, ou mesmo de lugares mistos, ele está prejudicando a inteireza do corpus e isso vai impactar nos resultados alcançados. Porque trata-se de uma forma artificial de manipular o conjunto de palavras.

O que eu descrevo no parágrafo acima, pode ser facilmente percebido submetendo um corpus inteiro à análise pelo Lexico3 e depois **rodando** o mesmo corpus com algumas dezenas de balizadores[11]. As diferenças saltam aos olhos quando os gráficos são apresentados. E então fica claro que a estratégia de subtrair formas de um corpus para que ele fique do mesmo tamanho do outro é completamente viciosa.

Para evitar a intromissão humana e poder comparar textos de tamanhos desiguais, alguns pesquisadores desenvolveram fórmulas para minimizar os erros possíveis na comparação de corpora de tamanhos vários. E veremos a seguir que esse é um problema antigo dos linguistas. Mesmo a fórmula mais simples, citada acima, data de 1957.

Ao desenhar a arquitetura de NEOLO, eu previ inserir ferramentas que dessem conta de fazer os cálculos de riqueza lexical de corpora de tamanhos desiguais para análises comparativas. Após muita leitura e contato com pesquisadores da área e com estatísticos, decidi por 3 (três) procedimentos: MSTTR (Mean Segmental Type-Token Ratio; Johnson, 1944); MTLD (Measure of Textual Lexical Diversity; McCarthy 2005); e, HD-D - os autores não nomearam esse acrônimo, mas subentende-se que se trata de hypergeometric probability distribution com mais um D que aparece significando “resultado”.

MSTTR[12] - Neste processo, o texto a ser analisado é dividido em segmentos iguais em termos do número de palavras (normalmente 100 palavras por segmento). Para cada segmento, o

[11] As balizas no Lexico3 são partições inseridas pelo pesquisador no corpus para verificar o comportamento de determinada qualidade nas várias partes do texto. Aí percebemos como certas características aparecem fortemente em algumas partes do texto e como não ocorrem em outras partes, ou ocorrem modestamente em outras.

[12] Neste ponto em que explico as três ferramentas de cálculo da RL de textos (corpora), parafraseei os textos de Torruella e Capsada, 2013 e McCarthy, 2010. Justifico que procedi assim para evitar tratar dos conceitos livremente e cometer algum erro que prejudique a compreensão das ferramentas.

TTR é calculado e usando uma média aritmética da TTR (Type/Token Ratio, mencionado mais acima neste texto) para cada segmento do MSTTR é obtido.

Observe que no processo descrito acima, não existe escolha das palavras que vão entrar em cada um dos grupos de 100 (cem) palavras. Elas são obtidas por segmentos de uma centena de palavras. Em nosso entendimento, essa fórmula fica prejudicada pelo fato de diferentes partes dos textos terem características, às vezes, muito diferentes e, em assim sendo, vai gerar TTR que não são compatíveis com a média dos TTRs do texto. Para entender ao que me refiro, retome a leitura quando falo dos balizamentos do Lexico3.

MLTD -Este índice é semelhante ao do MSTTR, pois também opera dividindo o texto em segmentos e o TTR é calculado para cada um. Mas, neste caso, o comprimento dos segmentos é variável e depende do valor alcançado pelo TTR do segmento, isto repetido até o final do texto. Cada segmento termina quando seu TTR atinge um valor de 0,72. Ao final do texto é aplicado o cálculo MLTD = L / n, onde L é comprimento do texto em número de palavras e n é o número de segmentos. Novamente, minha crítica cai sobre o fato dos conjuntos de palavras serem segmentos de texto, então caí no mesmo vício do MSTTR.

Chamo a atenção que os erros que podem aparecer pelo fato de as palavras serem escolhidas atendendo os segmentos pode ser nulo, ou quase nulo, mas em determinados textos essa diferenção pode ser alta. Especiamente em textos em que o TTR geral é muito alto (e. g. *Catatau* de Paulo Leminski).

HD-D – Esta é a fórmula para desobrir a riqueza lexical de textos com número de formas diferentes mais complexa. Ela leva vantagem sobre as demais porque não compõe os grupos de palavras para serem analisadas sem ser segmentos. Com isso evita-se o problema que levantei nos 2 (dois) parágrafos acima.

Dentre as 3 (três) fórmulas que encorporei a NEOLO, essa é a mais complexa e a que deu mais trabalho para ser compreendida e desenvolvida.

A ferramenta funciona, grosso modo, dessa forma: ela forma grupos de 42 (quarenta e duas) formas retirada aleatoriamente das diversas partes do texto. Importante que eu diga: essas 42 (quarenta e duas) palavras não voltam para o texto. Então, para cada tipo lexical, calcula-se a

probabilidade de ele ser encontrado nas 42 (quarenta e duas) formas selecionadas randonicamente. A probabilidade para cada tipo lexical é somada e o resultado é usado como um índice para determinar a riqueza lexical. Este procedimento é conhecido dos estatísticos como uma distribuição hipergeográfica.

Cabe aqui uma consideração importante sobre a HD-D: ela é muito demorada, mesmo em computadores rápidos e eu não estou aqui me referindo a PC. Relato que quando a ferramenta foi testada, para rodar o texto Ulysses de James Joyce gastei 5 (cinco) horas utilizando o *mainframe* da UW-Seattle.

Por conta desse tempo utilizado para se ter o resultado da HD-D. Ou utilize textos curtos, ou utilize estações de trabalho com processadores Microsoft/Xeon. Estes *chips* não esquentam e podem passar muitas horas ligados sem serem danificados, além de serem bastante rápidos.

6 PALAVRAS FINAIS

A narrativa que foi desenvolvida até aqui deu conta de um estudo de caso do qual resultou o software NEOLO que embora já esteja 100% funcional, mas ele está em permanente mudança. No último ano, o Dr. Ariel colaborou muito com o fornecimento dos programas de *stemming* e, quiça, ele consiga colocar o programa para uso diretamente on-line. Este seri um passo muito importante por disponibilizar o programa para pessoas que não queiram, ou não possam, manusear o soft em ambiente LINUX. Além dos acréscimos de funcionalidades, existem mudanças no aparato de NEOLO quase diariamente. Os bancos de palavras são exemplos desses anexos ao NEOLO que estão em constante acréscimo. Isto acontece todas as vezes que encontramos palavras estranhas ao bancos.

Com isso concluimos a história de NEOLO, este que vem nos rendendo pesquisas interessantes e promete ser muito útil para o grupo de trabalho em que me integro.

REFERÊNCIAS

ALUÍSIO, S. M.; Gladis ALMEIDA. O que é e como se constrói um corpus. *Calidoscópio*, v. 4, n. 3, p. 156-178, set./dez. 2006.

CALABRESE, Omar. *Neo-Baroque: A Sign of the Times*. New Jersey: Princeton U. P., 1992.

FIGUEIREDO, Cândido de. *Novo Dicionário da Língua Portuguesa*. 1913.

FREDRIC, Jameson. *Postmodernism, or, the cultural logic of late capitalism*. Durham: Duke U. P., 1999.

GALEANO, Eduardo. *As veias abertas da América Latina*. Rio de Janeiro: Paz e Guerra, 1971.

HUTCHEON, Linda. *A poetics of Postmodernism*. London: Routledge, 1992.

LAWRENCE, Graham. *Webster's Unabridged Dictionary*. Disponível em: <http://unabridged.merriam-webster.com/>. Acesso em: 20 jan. 2017.

LEMINSKY, Paulo. *Catatau*. São Paulo: Iluminuras, 2000.

McCATHY, Philip M.; MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, v. 42, n. 2, p. 381-392, 2010.

MONSIVÁIS, Carlos. The Neobaroque and popular culture. *PMLA*, v. 124, n. 1, p. 181-188, 2009.

PINTO, Luiz M. da S. *Dicionário Brasileiro da Língua Brasileira*. Ouro Preto: Tipografia do Silva, 1832.

SARDUY, Severo. *Barroco*. Paris: Gallimard, 1975.

SARDUY. Severo. The Baroque and the Neobarque. In: KAUP, Monika; ZAMORA, Lois. (Eds.) *Baroque new worlds: representation, transculturation, countercoquest*. Durham: Ducke U. P., 2010.

TAVARES, Camilo de. *Um ano que durou 21*. Documentário. 2013.

TORRUELLA, J. e Ramon CAPSADA, Lexical Statistics and Tipological Structures: A Measure of Lexical Richness. *Procedia*, n. 95, p. 447-453, 2003.

WACKER, Kelly A. Baroque tendencies in contemporary art. In: _____. (Ed.). *Baroque tendencies in contemporary arts*. Cambridge: Cambridge Scholars Publishing, 2007.

Recebido: 17 de julho de 2017
Aceito: 19 de julho de 2017

ANEXO

Sintaxe para rodar NEOLO

$> ./neolo Texts/Ulysses.txt --mltd --msttr --hdd -t -w -p -v -x --dicts/Webster.txt listas/Eng385.txt nomes/20names8.txt -s abbrev/eng.txt> ~/Desktop/Ulyssestest.txt

Opções que podem ser acrescidos ou retirados dependendo da necessidade do pesquisador

-h, --help show this help message and exit

./neolo roda o programa NEOLO

--dicts DICT [DICT ...] a list of reference texts to compute neologism count

--mltd measure of lexical textual diversity

--msttr mean segmental type-token ratio

--hdd HD-D probabilistic TTR

--verbose, -v increase the verbosty (can be repeated: -vvv)

--wordlen, -w print the distribution of words by length

--wordtypes, -t print the distribution of wordtypes (unigrams) by count

--hapax, -x print the list of hapax legomena

--punc-ratio, -p print the ratio of punctuation tokens out of total tokens

--no-hyphen, -y remove the hyphen (-) from the list of punctuation symbols used in tokenization

--no-apostrophe, -a remove the apostrophe (') from the list of punctuation symbols used in tokenization

--sents [ABBREV], -s [ABBREV] print sentence length statistics, uses an (optional) abbreviations file containing stings which don't end sentences (eg: Mr.).

One abbreviation per line, include relevant punctuation.

Note that items in the abbreviations file will also be protected during later tokenization.

Select one or more statistics via the above options.

Note that some options with optional args (--sents/-s, eg) or options which take lists of files (--dicts, eg) can create ambiguity parsing the command line. When using these options, it may be best to place them after the name of the text you're trying to analyze:

$ ./neolo mytext.txt --dicts d1.txt d2.txt -sents abbrevs.txt

HD-D is a very sophisticated statistical measure of text richness for this it takes many hours to give results (e.g. to analyze Ulysses, the main frame computer at UW took 5 hours to process

Ulysses - Joyce). It is recommended to use HD-D only in Xeon Machines Family. Or in analyze of short texts.