<REVISTA TEXTO DIGITAL>

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

ESTUDOS ESTATÍSTICOS DE TEXTOS LITERÁRIOS

STATISTICAL STUDIES OF LITERARY TEXTS

Verônica Ribas Cúrcio
Graduada em Letras pela UFSC
Universidade Federal de Santa Catarina
demodoco@gmail.com

RESUMO: Neste texto apresentamos um recorte do que já foi comumente desenvolvido no âmbito de estudos estatísticos textuais, apontando alguns programas criados para esse tipo de tarefa e introduzindo a nomenclatura corrente de tal pesquisa.

PALAVRAS-CHAVE: estatística textual, análise quantitativa, tratamento informático de textos

ABSTRACT: In this text we present a summary of what has been developed in the area of textual statistical studies, pointing out some programs created for this type of task and introducing the current nomenclature of such research.

KEYWORDS: textual statistics, quantitative analysis, computing treatment of texts

As análises quantitativas e estatísticas de textos literários são, geralmente, acompanhadas por algum programa criado para auxiliar, a pesquisa e a exploração dos textos (em âmbito literário, lingüístico, sociológico, entre outros.). Realizações como contagem de palavras, distância lexical, crescimento de vocabulário, bem como a sua riqueza, e tantas outras características textuais, são possíveis através de uma rápida manipulação informática. Se tais realizações fossem feitas manualmente, demandariam muito mais tempo inviabilizando a pesquisa. Queremos apresentar a aceleração que essa ferramenta dá ao estudo, de modo que após isso, possa ser feita a parte que cabe somente ao pesquisador, a de analisar e

interpretar os dados que lhe aparecem na tela do computador. Lembramos também que esse processo é constituído por duas partes conjuntas: o trabalho do pesquisador, que é submetido às regras informatizadas, e o programa a ser utilizado, que vai necessitar de diretrizes que serão organizadas tanto pelo programador quanto pelo pesquisador, para que assim o programa possa trabalhar de modo satisfatório. A soma dessas duas partes resulta em uma parceria entre os técnicos que desenvolvem e aprimoram o programa, e os pesquisadores da área das ciências humanas que irão utilizá-lo, verificando a melhor maneira de se investigar o objeto textual. É um exercício mútuo e dessa união já resultaram muitos aplicativos[1]. Através de tais programas (desde que estejam alimentados por textos que serão alvo de estudos), podemos obter a riqueza lexical de um autor, verificar a abrangência de seu vocabulário, e, a partir de análises, ainda atribuir a algum texto apócrifo a sua autoria, chegando mesmo a realizar estudos temáticos ou estilísticos. Enfim, realizar estudos que não são de práticas recentes, porém, passíveis de quantificação. Certamente, nem todas as informações que os estudos de tratamento estatístico trarão serão novas. Contudo, é uma outra maneira, talvez um pouco mais objetiva e veloz, de dizer o que outrora foi discutido, de afirmar com mais segurança ou até mesmo ampliar o horizonte de algumas certezas já enunciadas.

Michel Bernard[2] (1999), por exemplo, protesta contra o uso da expressão "novas tecnologias" pelo fato de o computador já existir há mais de meio século e de os pesquisadores em literatura utilizarem-se também de ferramentas informáticas. Ele aponta que um grande número de pesquisadores e estudantes consulta, com regularidade significante, as bibliografias e enciclopédias eletrônicas que estão à disposição nas bibliotecas universitárias, lembrando também do uso de correio eletrônico e sítios criados cada vez mais, em maior quantidade. Ele mostra como a informática é utilizada para a pesquisa e qual é a sua eficácia, afirmando que há mais de trinta anos os computadores são utilizados para a realização de estudos no campo da literatura, já existindo uma enorme quantidade de

trabalhos publicados acerca deste assunto. Para ele, os trabalhos informatizados apenas dão continuidade, através dessas ferramentas, às tradições seculares de pesquisa nos campos filológico e literário. A exemplo disso, lembra que, desde a Idade Média, já era praticada, de uma forma manual, a estatística textual como a busca de concordâncias[3] e a indexação. Tudo isso seria apenas um reencontro, bastante natural, entre as técnicas de pesquisa literária e as ferramentas que facilitam uma grande empreitada.

Bernard posiciona como primeiro problema para a realização desses estudos o custo dessa tarefa. Pois até então, a pesquisa literária era a menos custosa em material: "Une bibliothèque (...), du papier, une machine à écrire, et voilá le littéraire équipé pour les études les plus poussés." (BERNARD, 1999: 9) O autor mostra que a realização da pesquisa literária exigia pouco custo, pois o pesquisador podia realizar sua tarefa em casa, favorecendo, por conseqüência disso, uma pesquisa individualista ou solitária. O único custo se baseava no tempo de leitura e no tempo de escrita. Com a introdução da informática, os custos mudam esse panorama. As máquinas, os programas, os tempos de cálculo, os técnicos, as salas especializadas, as horas de conexão (hoje temos outros meios de conexão que também trazem outro tipo de investimento) têm um custo significativo, que acarreta uma reavaliação dos objetivos da pesquisa. Por outro lado, essa mudança também invoca a necessidade de cooperação: ao contrário da idéia anterior de pesquisa individualista, ela exige um laço imprescindível entre pesquisadores de diferentes áreas, como as ciências exatas e as ciências humanas. Bernard ressalta a economia no tempo de pesquisa, pelo uso do computador, facilitando o percurso para a continuidade da mesma, bem como a redação dos resultados. Ou seja, o uso da informática impõe uma preparação (digitalização, tratamento e correção de textos, para que os mesmos possam ser submetidos a uma análise informatizada) mas, logo em seguida, permite retomar o tempo que se perdeu na execução dessa tarefa.

Conforme Bernard, a ferramenta da informática permitiu o surgimento de novas categorias de análise de textos. Por exemplo, em uma análise de um dado corpus, pode-se afirmar com segurança, que um termo é ausente. Tal reconhecimento não poderia ser feito com certeza absoluta, se realizado manualmente. Outra possibilidade é a caracterização de uma obra, sabendo através do seu vocabulário, não somente o que ela diz, mas também o que ela evita dizer.

Bernard ressalta que para realizar uma melhor leitura dos resultados de um tratamento lexicométrico, é bom lembrar que os computadores têm apenas acesso a códigos numéricos e jamais ao sentido das palavras, quer dizer, a máquina é (ainda) incapaz de compreender informações semânticas ou simbólicas.

Para o autor, a primeira operação executada pelos programas específicos em análise textual é um corte das seqüências de caracteres em *formas*, de onde eles retiram as *ocorrências*. Por exemplo, se a forma "garrafa" aparece uma vez em um dado corpus, se diz que ela tem uma *ocorrência*, e que tal é a sua *freqüência*. O computador, mais especificamente, o programa é capaz de distinguir e fornecer uma lista (index ou índice) de todas as formas do texto escolhido, em ordem alfabética ou em ordem decrescente de freqüências. Geralmente as formas mais freqüentes na listagem são as chamadas *palavras gramaticais*, também encontradas como *palavras funcionais*[4], como os artigos, os dêiticos, as preposições, os pronomes, as conjunções. Para Bernard, essas palavras podem trazer indicações estilísticas, porém, são menos significantes que os primeiros substantivos listados, pois os mesmos carregam o peso temático. O autor lembra que, salvo os textos lematizados (veremos mais adiante), os verbos aparecem com menor freqüência porque suas formas são muito variadas, diferente dos substantivos, e por conseqüência disso, os verbos acabam se espalhando muito mais pela listagem.

Ludovic Lebart e André Salem[5] (1994: 15) afirmam que os sucessos obtidos pelas aplicações do método estatístico em muitos domínios, como na física ou na biologia, ganharam espaço

nas ciências humanas compreendendo as disciplinas que se utilizam da linguagem, principalmente em estudos de vocabulário. Os primeiros trabalhos realizados nesse âmbito com o método estatístico eram voltados para a estatística lexical. Para os autores, essa corrente de estudos se fixou nos objetivos e preocupações que são anteriores a aparição dos métodos quantitativos, como medidas comparativas do vocabulário de diferentes autores, medida da evolução do vocabulário de um mesmo autor durante um período de sua produção, entre outros. Paralelamente a isso, outro método que se desenvolveu foi a lingüística estatística, que segundo Herdan seria "la quantification de la théorie saussurienne du language" (HERDAN apud LEBART, 1994: 17). Ou seja, seria uma ramificação da lingüística estrutural, tendo como principal função a descrição estatística do funcionamento das unidades definidas pelo lingüista aos diferentes níveis de análise lingüística.

Uma amostra histórica

Anthony Kenny (1982:1) apresenta um recorte histórico sobre trabalhos que já foram desenvolvidos neste campo. Ele afirma que o início dos estudos estatísticos de estilo nos tempos modernos é comumente relacionado ao ano de 1851, quando Augustus de Morgan contestou a autenticidade de alguns escritos do apóstolo Paulo. A idéia era estabelecer a autoria através da medida do comprimento de palavras, ou seja, do número de letras por palavras usadas nas várias Epístolas. Segundo Kenny, a primeira pessoa que de fato testou as hipóteses do comprimento de palavras para verificar uma característica distinta de escritores foi um pesquisador norte-americano chamado T. C. Mendenhall. Sua tentativa mais ambiciosa nessa direção foi um estudo da controvérsia Shakespeare-Bacon no *Popular Science Monthly* em 1901. Através de uma doação de um filantropo de Boston, Mendenhall conseguiu empregar duas secretárias e uma máquina calculadora para analisar cerca de 400.000 palavras de Shakespeare, 200.000 palavras de Bacon e quantidades de textos de autores de outros períodos. Mendenhall descobriu que o vocabulário de Shakespeare consistia em palavras cuja média do

comprimento ficava abaixo de quatro letras, menos que qualquer escritor de língua inglesa antes estudado; e que sua palavra de maior freqüência foi caracterizada com quatro letras, uma coisa jamais vista anteriormente. Essas características destacaram Shakespeare da maior parte de seus contemporâneos, bem como de autores do século XIX previamente estudados. Kenny aponta que enquanto Mendenhall usava estatísticas estilísticas na tentativa de resolver problemas de atribuição em inglês, escolas européias desenvolviam técnicas estilométricas para a língua grega a fim de estabelecer o problema cronológico dos diálogos de Platão. Conforme Kenny, Lewis Campbell ofereceu sua edição de 1867 de *Sofista* e *Política* à uma bateria de testes estilísticos, mostrando que Campbell acreditava serem indicadores de dados relativos: ordem de palavras, ritmo, medidas através de freqüência de *hapax legomena* (palavras de uma única ocorrência) e principalmente 'originalidade de vocabulário'. Campbell concluiu que o *Sofista* e *Política* pertencem a um período tardio da vida de Platão, ou seja, depois de *A República* e próximo a *Leis*. O autor diz que o trabalho de Lewis Campbell passou despercebido durante aproximadamente 30 anos, até que um filólogo alemão denominado Ritter, em 1888, obteve conclusões semelhantes através dos mesmos métodos. Outra pesquisa que Kenny nos traz baseada em estudos estatísticos diz respeito ao trabalho de Udny Yule. O mesmo realizou uma comparação estatística entre trabalhos de Jean Gerson e Thomas à Kempis para buscar a autoria do texto *Imitatio Christi*. Desses estudos, que transcorreram um percurso de 1938 a 1944, resultou uma publicação denominada *The statistical study of literary vocabulary*, obra pioneira nesse campo.

Kenny nos afirma que os estudos relacionados à estatística de textos se desenvolveram enormemente entre Mendenhall e Yule e que, com a invenção do computador, surgiram grandes pesquisas quantitativas em estudos de textos literários. Em seguida, Kenny ilustra alguns tipos de problemas estudados com esse remanejamento de ferramentas e os diferentes métodos utilizados para esse novo tipo de empreitada. Os estudos sobre o Novo

Testamento, tomando como ponto de partida a questão da autoria das Epístolas, foram novamente recapturados por W. C. Wake e A. Q. Morton. Ambos estudaram o comprimento das frases nas Espítolas e fizeram extensas comparações entre esse mesmo comprimento com outros autores gregos.

Para o início dos anos sessenta, Kenny aponta dois estudos de textos em língua inglesa que são considerados como modelos de pesquisa estatística de estilo, um dentre eles é o estudo do historiador literário sueco Alvar Ellegård sobre *Junius Letters* (cuja autoria era desconhecida). O historiador dirigiu seu estudo com a nova pesquisa de estatística lingüística, apresentando uma lista de aproximadamente quinhentas palavras e expressões que caracterizavam a escrita de Junius. Chegou à conclusão que Sir Philip Francis era Junius. O outro trabalho apontado por Kenny é de Frederick Mosteller e David Wallace sobre *Federalist Papers*. Os autores usaram técnicas mais elaboradas que as apresentadas pelos trabalhos de Ellegård, com métodos estatísticos baseados nos teoremas de Bayes[6].

Uma amostra do que se faz atualmente

Outros pesquisadores que também desenvolvem estudos na área de estatística textual são os franceses Lebart e Salem. Para eles, a estatística textual surgiu a partir do encontro de interesses de duas áreas: a lingüística e a estatística. Desse encontro, outros campos de estudos também usufruíram significativamente, como a análise do discurso, análise de conteúdo, pesquisa documentária, etc. (LEBART, L.; SALEM, A., 1994:11) Segundo os autores, também a estatística textual empreende resolver uma série de problemas levantados por estilísticos em seus estudos de vocabulário de autores, algumas preocupações que já eram significativas antes mesmo da aparição das pesquisas quantitativas, como por exemplo, medidas comparativas de vocabulário de diferentes autores, medida da evolução do vocabulário de um mesmo autor ao curso do período enquanto foram produzidas suas obras, entre outras. Tais métodos de estatística textual podem objetivar ainda mais as contagens

anteriores, intuitivamente contribuídas pelos próprios estilísticos, podendo acentuar tais trabalhos ou até mesmo refutá-los:

> Chaque fois qu'on se risque à dire d'un auteur qu'il aime, ou qu'il préfrère une tournure, un thème, un effet de style, chaque fois qu'on utilise pour le caractériser les mots fréquent, rare, souvent, jamais, même, autre, tout, recherché, banal, comum, original, caractéristique, typique, etc..., on fait appel à une statistique implicite, à des fréquences attendues et observées et en fin de compte à la notion d'écart. (BRUNET, 1983:8)

Lebart e Salem (1994:19) apresentam didaticamente o procedimento que ocorre na maior parte das aplicações estatísticas, organizando a pesquisa em uma seqüência de quatro fases. Para os autores, cada uma dessas fases poderá sofrer qualquer implicação segundo o contexto, as aplicações e os domínios a serem praticados:

1. A primeira fase se constituiria com o *problema*, ou seja, aquilo que motivou o estudo.
2. Os *dados* podem ser de natureza experimental ou provindos de alguma observação.
3. O *tratamento* é entendido como a fase constituída, no caso inferencial mais clássico, para colocar em provas as hipóteses, ou mesmo modelos. Em casos descritivos ou exploratórios essa fase é sobretudo uma forma de aplicar os dados destinados a apresentar os traços estruturais de maior importância.
4. E finalmente, a *interpretação*, que seria a avaliação crítica das hipóteses ou do eventual modelo de partida. No caso descritivo ou exploratório, essa fase compreenderá inevitavelmente uma reflexão sobre a validade e a significação das estruturas observadas.

Lebart e Salem (1994:33) retratam a necessidade que pesquisadores, em diversos domínios científicos, têm de comparar textos sobre bases quantitativas. A experiência do tratamento lexicométrico de conjuntos textuais reunidos a partir de problemáticas diferentes mostra que, mediante uma

adaptação mínima, um mesmo conjunto de métodos encontra aplicações pertinentes em numerosos estudos de caráter textual. Em cada caso, o recurso aos métodos quantitativos é motivado por preocupações e objetivos distintos, por exemplo, estudos estilométricos comparados a textos de diferentes autores, tipologias de respostas de indivíduos a uma mesma questão aberta ou pesquisa documentária.

Conforme os autores, o método estatístico tem como alicerce medidas e contagens realizadas a partir de objetos que se deseja comparar. Esse tipo de procedimento requer que identifiquemos tais objetos com uma certa ordem, uma nomenclatura. Como as *ocorrências* idênticas de um mesmo tipo ou de uma mesma *forma*. Para submeter um conjunto de objetos a comparações estatísticas, devemos definir uma série de relações sistemáticas entre os casos particulares e as categorias mais gerais.

Na prática, Lebart e Salem afirmam que a aplicação desses princípios gerais implica a definição de uma norma que venha permitir o isolamento das diferentes unidades de uma seqüência textual. A operação que permite recortar o texto, ou o conjunto de textos, em *unidades mínimas* (unidades que serão decompostas mais adiante no procedimento estatístico) é chamada de *segmentação* do texto. Logo após esse recorte, sucede uma próxima etapa, a fase de *identificação* onde ocorre uma outra fragmentação do texto em unidades distintas, ou melhor, em um reagrupamento, a partir do texto, das unidades idênticas.

Conforme os autores (LEBART, L.; SALEM, A., 1994:36), para que a realização de uma segmentação automática do texto em ocorrências de formas gráficas se concretize, é necessária a escolha, por meio de um conjunto de caracteres, de um subconjunto que se designará sob o nome de *caracteres delimitadores*. Uma seqüência de caracteres não-delimitadores limitada às suas duas extremidades por caracteres delimitadores é uma *ocorrência*, duas seqüências de caracteres não-delimitadores constituem duas ocorrências de uma mesma *forma*. E

o conjunto de formas de um texto constitui o seu *vocabulário*. A segmentação assim definida permite considerar o texto como uma seqüência de ocorrências separadas entre elas por um ou mais caracteres delimitadores. O número total de ocorrências contidas em um texto é o seu *tamanho* ou seu *comprimento*.

É possível operar reorganizações das formas e das ocorrências do texto: as ocorrências de uma mesma forma se encontrarão agrupadas em uma mesma direção, acompanhadas de um pequeno fragmento do contexto imediato, onde se fixará o comprimento em função de necessidades particulares. Chamamos *forma-pólo* a forma onde se reagruparão os contextos. Esse tipo de reorganização permite estudar mais facilmente as relações que podem existir entre os diferentes contextos de uma mesma forma.

A distinção de caracteres do texto em caracteres delimitadores e não-delimitadores permite definir uma série de descritores relativos às formas simples. Para abordar a descrição dos segmentos compostos de várias formas e repetidos em um corpus de textos, é necessário especificar o estatuto de cada um dos caracteres delimitadores.

Para os procedimentos formais que comumente se elaboram, nos contentaremos em dar a alguns signos de pontuação (o ponto final, o ponto de exclamação, o ponto de interrogação) o estatuto de separador forte ou separador de frase. Entre esses caracteres delimitadores nós escolheremos igualmente um subconjunto correspondente às pontuações fracas e fortes (em geral: a vírgula, o ponto e vírgula, os dois pontos, as aspas e os parênteses) e chamaremos o conjunto de delimitadores de *seqüência*. A continuação, então, das ocorrências situadas entre dois delimitadores de série é considerada como uma *seqüência*.

Lebart e Salem (1994:36) declaram que, privilegiando o ponto de vista lexicométrico, podemos, em algumas situações, considerar indispensável antes de todo o tratamento quantitativo sobre o corpus, submeter as unidades gráficas oriundas da segmentação automática a uma *lematização*. Ou seja, dar regras de

identificação que permitem reagrupar nas mesmas unidades as formas gráficas que correspondam às diferentes flexões de um mesmo lema. Os autores exemplificam o modo convencionado de lematizar o vocabulário de um texto em francês. Tal modo leva geralmente:

- as formas verbais ao infinitivo
- os substantivos ao singular
- os adjetivos ao masculino singular
- as formas ligadas, sem a ligação

A lematização é um processo que atua como uma espécie de filtro, deixando o número total de formas de uma maneira que evite a sua repetição. Como, por exemplo, as diferentes conjugações de um mesmo verbo, ou as flexões que indicam número e gênero, etc. A lematização é uma maneira de garimpar o texto, de modo que as formas sejam contabilizadas mais estritamente.

Através da busca de palavras (formas) podemos verificar o seu modo de emprego e quais as palavras que se encontram vinculadas a ela em maior ou menor grau. A busca, dependendo do programa, pode proporcionar um raio de palavras tanto à esquerda da forma-pólo, quanto à sua direita. Com isso podemos buscar as *co-ocorrências*, ou seja, as palavras que surgem normalmente na mesma sentença, no mesmo parágrafo, ou em um mesmo contexto.
Existe uma maneira de verificar a freqüência das ocorrências e suas localizações, esse modo de verificação geralmente é designado como *índice*. Ele é um sistema de organização, onde se apresentam todas as formas, e dependendo do programa, essas formas e suas ocorrências podem vir acompanhadas de sua freqüência e localização no corpus. Para encontrar tais ocorrências no texto de partida, utiliza-se de preferência um sistema de endereços, um sistema de coordenadas numéricas mais estritamente ligadas à edição do texto como: o tomo, a página, a linha, a posição de ocorrência na linha, etc. Essas informações, que permitem retornar mais facilmente ao documento de origem, são as *referências* associadas a cada uma das ocorrências. Os índices podem classificar as formas segundo

critérios diferentes. Chamamos de *índice alfabético* o índice no qual as formas são classificadas segundo a ordem lexicográfica, ou seja, a ordem corrente dos dicionários; e chamamos *índice hierárquico* aquele no qual as formas são posicionadas em ordem decrescente, segundo as suas freqüências.

Para os autores (LEBART, L.; SALEM, A., 1994:55), em um corpus de tamanho grande, o seu crescimento do vocabulário tende a sofrer uma dupla influência:

- A apreensão de novas ocorrências tende a aumentar o número total de formas de um corpus (mais ocorrências, mais formas distintas).
- Quando o tamanho do corpus aumenta, a taxa de formas novas trazidas para cada crescimento do número de ocorrências tende a diminuir.

Charles Muller, em *Initiation a la statistique linguistique*, nos explica de uma outra maneira o crescimento do vocabulário. Passo a passo, Muller descreve como analisar esse crescimento tendo um dado texto. Ele começa por contar as palavras que compõem esse texto, obtendo assim um valor numérico *N* (que é o número total de palavras); tal número será a medida de extensão do texto. Em seguida, o autor une cada uma dessas palavras a um vocábulo (forma), para obter um segundo valor numérico, *V*, (que designa o número de vocábulos que tem uma ocorrência, no mínimo, durante a extensão do texto). Para Muller, V está em função de N, ou seja, para um texto qualquer, V tende a crescer com N. Sendo evidente que V cresça mais lentamente que N, pois cada palavra que representa um vocábulo já utilizado no texto, atrasa o número de V em relação ao de N. Ao iniciar uma contagem de um texto, o autor repara que V é igual a N até a primeira repetição de um vocábulo qualquer.

Então, como visto logo acima, a extensão do vocabulário (V) está em função da extensão do texto (N). Muller segue com outro exemplo: de um texto qualquer considerado homogêneo, ele extrai dois fragmentos de comprimento distintos. Deve-se prever que o mais longo terá um vocabulário de extensão superior em relação

ao mais curto. Porém a extensão do vocabulário está em função também do estilo, ou seja, ele é determinado, no mínimo, pelo léxico do autor na situação estilística onde ele se encontra. Muller diz que se recolhermos dois textos de estilos muito distintos e de comprimento igual, observaremos um desvio entre a extensão do vocabulário dos dois textos, e esse desvio é uma característica estilística de primeira importância; admitiremos assim que o vocabulário mais extenso significa também um léxico mais extenso. (MULLER, 1968: 156-157)

Em seu trabalho sobre o vocabulário de Proust, Etienne Brunet (1983: 20) dá sua explicação a respeito do crescimento de vocabulário, afirmando ser uma noção relativa e dinâmica, ao contrário da riqueza lexical, que se apresenta como uma medida absoluta, independente da ordem dos textos considerados. Ao colocar os sete textos proustianos em ordem cronológica, Brunet percebe que, na seqüência dos textos, as entradas de palavras novas eram cada vez mais raras na medida em que a contagem ia chegando ao final da obra.

Isso aponta, mais uma vez, para a explicação de que quanto maior for o tamanho do corpus investigado, menor será o número de formas novas que ele irá apresentar.

Para Lebart e Salem (1994:246), muitos autores utilizam a noção de "palavras-funcionais" se apoiando sobre uma intuição comum que se pode reunir em cada língua, uma lista de formas que se costuma chamar, às vezes, de "palavras-gramaticais" e que têm em comum a propriedade de ser menos marcadas ao plano semântico. Os autores indicam um trabalho desenvolvido por Demonet et al. (DEMONET apud LEBART, 1994: 246) sobre essa tal lista de "palavras funcionais": eles propuseram medir uma "taxa de funcionalidade" própria a cada discurso, ou a cada tipo de discurso, recenseando o número de ocorrências que correspondem às ocorrências de uma lista de "palavras funcionais" previamente listadas por eles. Essa propriedade tem sido amplamente utilizada, nos estudos informatizados, para separar as "palavras funcionais" da lista das formas consideradas

unidas, correspondendo em grande parte, às formas mais freqüentes, reputadas como pouco dignas de interesse.

O trabalho do pioneiro Ellegård (1962), que citamos no início do texto, compara as proporções de "palavras-funcionais" no corpus de *Junius Letters* (panfletos publicados ao final do século XVIII, comportando cerca de 150.000 ocorrências), com um outro corpus da mesma época. Tal procedimento constitui uma fase importante dos trabalhos de Mosteller e Wallace (apud LEBART, 1994:246) e de Holmes (apud LEBART, 1994:246). Benzécri realizou tipologias de textos em grego antigo, latim e espanhol a partir de conjuntos de palavras funcionais, pondo em evidência, ao mesmo tempo, problemas que posam a seleção dessas unidades estatísticas, e o poder discriminante dos perfis de palavras funcionais quando essas intervem como elementos ativos de uma análise de correspondências (BENZÉCRI, apud LEBART, 1994: 246).

CONSIDERAÇÕES FINAIS

A idéia de fazer um apanhado histórico de estudos quantitativos em busca de estilo literário, autoria, vocabulário, entre outros, pretendeu apontar, de maneira sucinta, as possibilidades de pesquisa literária em parceria com a informática. Acreditamos que os estudos da literatura realizados em uma perspectiva quantitativa ainda têm muito a oferecer nacionalmente. Contudo, devemos estar atentos aos efeitos "ludibriantes" que os programas podem nos oferecer; é preciso saber o que se busca e interpretar o que se encontra, pois as possibilidades de ferramentas oferecidas são imensas, garantindo uma grande polivalência na pesquisa. O risco que corremos é de não haver aprofundamento nos dados levantados, resumindo o trabalho em apenas ilustrações de gráficos, de tabelas, com a presença de grandes números, algoritmos e cálculos sofisticadíssimos sem o arribo a interpretações consistentes que propiciem ao estudo legitimidade.

REFERÊNCIA BIBLIOGRÁFICA

BERNARD, Michel. *Introduction aux études littéraires assistées par ordinateur*. 1ª edition. Presses Universitaires de France. Paris, 1999.

BRUNET, Etienne. *Le vocabulaire de Proust I.Étude quantitative.* Slatkine: Genève, ?

KENNY, Anthony. *The computation of style. Un introduction to statistics for students of literature and humanities*. Pergamon Press: Oxford and New York, 1982.

LEBART, L.; SALEM, A. *Statistique textuelle*. Dunod: Paris, 1994.

MULLER, Charles. *Initiation à la Statistique Linguistique*. Librairie Larousse: Paris, 1968.

<TEXTO DIGITAL>

---

[1] *Lexico3*, *Hyperbase*, *WordSmith*, *TACT* (Text Analysis Computing Tools), *Pistes*, *Alceste*, *WordCruncher*, *Oxford Concordance Program*, *Saint-Clef*, *Le Concordeur*, *COMMAS II*, *FRECONWIN* e muitos outros

[2] Michel Bernard é integrante do grupo de professores pesquisadores do *Centre de Recherches Hubert-de-Phalèse*, que tem o intuito de difundir os métodos informáticos nos estudos da literatura. Publicam anualmente, desde 1991, trabalhos realizados através de ferramentas informáticas nos estudos literários.

[3] A concordância é uma forma de análise quantitativa de corpus que representa uma lista de ocorrências de uma ou de várias formas que fazem parte de um contexto do corpus.

[4] Em francês, são conhecidas como *mots-outils*, *formes fonctionnelles* ou *formes vides*. Abordaremos mais esses tipos de palavras mais adiante.

[5] Ludovic Lebart é diretor de Pesquisa do CNRS (Centre National de la Recherche Scientifique), da École nationale supérieure de Télécomunications e André Salem é engenheiro da École normale supérieure de Fontenay-Saint-Cloud.

[6] Podemos encontrar o artigo de onde deriva o teorema de Bayes no seguinte endereço: http://publicacoes.gene.com.br/ciencia_hoje/Bayes.pdf