LEXICOMETRY AND DISCOURSE ANALYSYS

LEXICOMETRIA E ANÁLISE DO DISCURSO

Dirceu Cleber CONDE
Universidade Federal de São Carlos (UFSCar)

RESUMO

Nosso intuito neste texto é demonstrar como um sistema de análise estatística lexical sobre corpora pode ser um aliado para a Análise do Discurso. Tal abordagem se baseia em conceitos teóricos consagrados, mas que não deixam de observar o fenômeno e suas características materiais. A Lexicometria toma parte nesse empreendimento não como um fim, mas como um meio auxiliar à reflexão do analista. Para tanto, apresento alguns exemplos a partir do sistema de análise lexicométrica denominado Lexico3.

ABSTRACT

Our aim in this paper is to demonstrate how a system of lexical statistical analysis of corpora can be an ally for doing Discourse Analysis. This approach is based on theoretical concepts enshrined, but we leave not observe the phenomenon and that material characteristics. The Lexicometry take part in this text not as an and, but as an support to work of analyst. For that, I show some examples from the lexicometry analysis system called Lexico3.

PALAVRAS-CHAVE

Lexicometria; Estatística; Análise do Discurso;

KEYWORDS

Lexicometry; Statistics; Discourse Analysis;

Introdução

Agradeço imensamente pelo convite para participar desta mesa-redonda[1]. Sinto-me bastante à vontade para falar de um lugar distanciado das grandes discussões da Análise do Discurso praticada no Brasil (AD), pois não milito mais nessa área e talvez, pelo fato de a minha opinião não ser a de um especialista é que ela não venha a despertar furor e controvérsia. Eu seria apenas mais uma voz de quem fala de fora da grita e que possa ser ouvida ou não. Mas o melhor de tudo é não ter compromisso com alguma corrente da AD e por isso nenhum compromisso de fidelidade teórica ou ideológica. Por isso vou me sentir muito à vontade para usar termos como "empírico", "estatística", "dados quantitativos", "recorrências quantificadas", "grupo de formas", entre outros. Essas primeiras palavras esquivadas da minha parte, no fundo, representam um pedido para que os analistas do discurso olhem para outras questões que na prática da análise foram obliteradas e se tornaram coisas de somenos. Tampouco tenho o direito dizer qual é a melhor ou a pior forma de se fazer. Meu objetivo é apenas demonstrar que uma possibilidade metodológica para a AD, dentre outras tantas possíveis, são os trabalhos de estatística lexical, em especial, a utilização de ferramentas de lexicometria, como uma entrada interpretativa para dados.

Como exemplo de ferramenta informatizada dedicada à análise lexicométrica, vou citar algumas operações possíveis com o software Lexico3 e dar sugestões para que futuras análises com corpus verbal possam ser aprimoradas e incrementadas através dessa tecnologia gratuita e acessível a todos interessados. Para tanto, este texto se organiza em duas partes: uma primeira que fundamenta o uso da máquina

[1] Texto apresentado por ocasião da minha participação na mesa-redonda intitulada **Discurso e "novos" diálogos teórico-metodológicos** no **V Colóquio da Associação Latino-americana de Estudos do Discurso – ALED BRASIL**, cujo tema fora **Análise de Discurso: novos canteiros de trabalho?** Realizado na Universidade Federal de São Carlos, 29 a 31 de maio de 2014. Meu especial agradecimento ao Professor Dr. Roberto Baronas pelo convite.

informatizada através da reflexão metodológica, conceituando alguns pontos sobre a lexicometria, a segunda parte apresenta uma singela de exploração de dados e possibilidades interpretativas para eles. Também desejo que minhas palavras sejam apenas de incentivo para novas (pelo menos aqui no Brasil) abordagens da materialidade do discurso. Por isso, esta apresentação não é um "manual" nem sequer publicidade de uma ferramenta, mas a abertura das inúmeras possibilidades das quais os analistas podem se valer no futuro.

1. O método a sua validade

Historicamente, a AD difundida, no Brasil é aquela baseada nos trabalhos de Michel Pêcheux. Há uma literatura bastante rica no País sobre a contribuição desse autor, levando o estado da arte a uma íntima identificação com o seu pensamento. No entanto, em nossa humilde opinião, a matriz desse legado chegou a nós de modo bastante particular e atualmente é desenvolvida por muitos pesquisadores através de modelos metodológicos bastante diversos ao proposto em sua origem. Novamente nossa opinião: quando uma teoria se torna muito "popular" seja pelo mérito de seus pesquisadores, seja por cair na graça de muitos, ela corre riscos de banalização, o que leva a algumas práticas que provocam um trocadilho um pouco perturbante, mas real: "os analistas do discurso precisam fazer mais análises e menos discurso".

Essa situação "um pouco perturbante" me leva a citar um texto de Pêcheux publicado em 1982 e traduzido para o português em 2011 como um exemplo bastante claro de que uma das principais referências da AD (no Brasil) não abandonou o sonho de um sistema informatizado e do tratamento de *corpora* diversificados e numerosos[2].

[2] Parece que essa ideia se coaduna bastante com a busca para se compreender como as "dispersões discursivas", neste caso, um conceito oriundo da leitura que se faz de Michel Foucault (2001), se comportam materialmente sob a forma de recorrências lexicais, sintagmáticas e outras formas materiais.

> A utilização da informática exige dos analistas de discurso uma construção explícita de seus procedimentos de descrição, o que é a pedra de toque da consistência de seus objetos teóricos. Ela permite, ainda, a apreensão de *corpora* variados de grande dimensão, o que consiste na pedra de toque da validade de seus objetos descritivos." (Pêcheux et Marandin, [1982]1990, p. 282.[3]

A analogia da "pedra de toque" é bastante interessante, pois esse instrumento é utilizado por ourives com o intuito de testar as ligas de metais precisos, ela é, normalmente um mineral lítico escuro rico em compostos silicosos. O teste com a "pedra de toque" consiste em fazer um risco como metal a ser testado, fazendo um risco de resíduo, (p.ex. uma aliança de ouro) e outro risco com o metal padrão (p. ex. uma ponta de ouro 18KT) sobre a pedra, em seguida, aplica-se o ácido adequado cuja composição reage com o ouro 18KT, se os riscos mantiverem a mesma coloração, significa que a amostra testada é ouro desse quilate, caso contrário o teste deve ser feito com outro padrão de quilate diferente até que se chegue à medida ou se conclua que a liga é falsa. A feliz analogia de Pecheux e Marandin nos apresenta um panorama que nos motiva a olhar para alguns aspectos e que estão organizados em duas dimensões: a) os procedimentos de descrição são a "pedra de toque" que comprova a consistência dos objetos teóricos; b) a quantidade é o que assegura a validade para seus objetos descritivos. Ora, o que há de peculiar nessa citação? Primeiramente o eixo "a" demonstra que para o analista de discurso não há um objeto teórico, pois o emprego está no plural, portanto, se para a AD seu objeto teórico é o Discurso, para o pesquisador, o(s) seu(s) objeto(s) teórico(s) pode(m) ser outro(s) objeto(s) sujeito(s) ao Discurso, portanto um pesquisador pode fazer

[3] O texto citado também se encontra traduzido integralmente para o portugês, ver Pecheux e Marindin (2011).

de uma manifestação discursiva um objeto teórico; no entanto o mais surpreendente está no eixo "b": esse(s) objeto(s) que o pesquisador elege só pode(m) ser testado(s) na "pedra de toque" da quantidade sobre os objetos descritivos, ou seja, da descrição de um dado fenômeno discursivo.

Ora, afinal, o que pode ser um objeto descritivo em AD? Acredito que há vasta bibliografia cuidando disso, mas vou arriscar meu palpite: os objetos descritivos de uma pesquisa em AD estão cravejados na materialidade linguística, nos valores diferenciais entre todos os níveis, desde o fonético-articulatório até as estruturas textuais mais complexas, passando pelas referências enunciativas e enuncivas das manifestações orais e escritas. Não seria uma modalidade anafórica usual/inusual um elemento indiciário de determinados posicionamento? A recorrência de terminado termo ou de conjunto de termos, com suas coocorrências, não seria algo para se rastrear em grupos de textos e de sujeitos enunciadores marcados social e historicamente? Determinada preferência para alguns termos em papéis temáticos de agente, paciente, beneficiário etc, não pode ser uma pista para um determinado posicionamento? Afinal, é a língua e sua estrutura que acionam não só o sentido como o efeito de sentido. Por isso vale lembrar Possenti (1996, p. 197):

> Tenho defendido sempre que posso que a língua é o material mais relevante do discurso, e que, portanto, uma AD deve ter uma boa e adequada teoria da língua, para extrair dela o máximo de proveito que puder.

Neste ponto, desejo adicionar mais um conceito que julgo ser bastante interessante a partir de uma outra característica sobre as línguas que pouco é lembrada por analistas do discurso, o fato de que "A Linguística é a ciência estatística tipo; os estatísticos sabem muito bem disso, a maioria dos linguistas ainda ignora tal fato.", (GUIRAUD, apud

LERBAT; SALEM, 1994, p. 18). Se a Linguística, enquanto ciência que se debruça sobre a língua, é prototipicamente uma ciência estatística ou pelo menos um espaço em que a estática pode render resultados, nada mais justo que olhar para a língua e seus fenômenos a partir das pistas recorrentes. Não se trata de defender uma teoria sobre a língua, mas um aspecto que julgo deveras relevante sobre as línguas: o de concordar com Guiraud e perceber que os fenômenos linguísticos também se comportam estatisticamente.

Olhar esse aspecto da língua não é um espaço muito confortável para o tipo de AD desenvolvida no Brasil, mas é bom lembrar que o levantamento de dados com a devida quantificação, organização e interpretação demonstra, de modo mais aprimorado, fenômenos que podem estar dispersos em diferentes manifestações textuais (orais ou escritas). Qual seria então a ligação entre dados estatísticos sobre uso de termos, sintagmas etc., e sua recorrência? É simples: podemos, por exemplo, entender que um determinado sintagma candidato à fórmula (CRIEG-PLANQUE, 2010) inicia sua jornada lentamente por algum grupo social ou se faz dele um jargão e quando menos se espera esse sintagma está em blogs, sites oficiais, textos jornalísticos, propagandas etc. Ora, esse crescente em ocorrência e disseminação pode ser um objeto teórico e pode ser rastreado, descrito e compreendido dentro de um panorama linguístico-textual (estrutura, composição, referencialidade, distribuição em sintagmas maiores, anáforas) que se liga ao efeito de sentido e explicita como determinadas relações do sentido podem estar ligadas às manifestações discursivas. Esse tipo de percurso retira o teor etéreo que às vezes se vê quando algum pesquisador fala sobre materialidade, mas não a demonstra, quantifica e qualifica nas análises de dados. Ou ainda, quando aplica interpretações sobre a materialidade que só se vê em poucos fragmentos de *corpus*. O que pouco corrobora com a perspectiva de que esse fenômeno possa surgir nos enunciados de mais e diversificados sujeitos.

Nesse caso, fundamentos estatísticos são de grande relevância para o tratamento dos dados, pois conceitos controvertidos, tais como "formação discursiva", "formação ideológica", "assujeitamento" podem, dependendo do método e de como se chegar aos dados, ser aplicáveis a grupos e populações de textos dispersos em diferentes categorias, tipos textuais, modos de manifestação ou sob o pretexto de qualquer outra divisão obviamente bem delimitada na metodologia. Por exemplo, um pesquisador poderia se dedicar a colher opiniões em blogs pessoais a partir de um tema comparando-as às opiniões de outros veículos como os textos institucionalizados em jornais e revistas sob o mesmo tema. Logo, os princípios estatísticos servem tanto para uma análise de textos (orais e/ou escritos) de um único autor, ou de um grupo de enunciadores, ou ainda de um período histórico, pois é bem provável que fórmulas, sintagmas, recorrências e distribuições de determinadas estruturas indiquem se um grupo compartilha ou não, por exemplo, de uma formação discursiva opondo-se à outra.

São muitas as pesquisas que se valeram ou se valem de alguma ferramenta informatizada para tratamento de dados lexicométricos em diferentes *corpora*. Não há espaço para apresentar tais trabalhos, tampouco esmiuçar seus matizes que são variados e ricos metodologicamente, por isso recomendamos a visita ao sítio eletrônico da revista Lexicometrica (http://lexicometrica.univ-paris3.fr) e também ao sítio do Centre de Lexicométrie et d'Analyse Automatique des Textes (SYLED-CLA2T) (http://syled.univ-paris3.fr/cla2t.html). Tratam-se de excelentes fontes sobre trabalhos realizados nessa área[4]. Esses vários trabalhos apresentam em comum, como ponto de partida para a explicação de fenômenos discursivos, os indicadores lexicométricos, levando em conta uma dada empiria muito salutar que parte dos dados para possíveis generalizações a respeito de algum fenômeno discurso, ou mesmo para a busca de dados negativos que frustram hipóteses, o que não deixa ter validade.

[4] Para discussões mais profundas a respeito de métodos estatísticos sobre o texto, recomendo a leitura de Lebart et Salem (1994), Benzécri (1981), Gilhaumou (1986, 1997).

2. O que é possível fazer com ferramentas de lexicometria

Há várias ferramentas de informática que podem ser adquiridas ou simplesmente baixadas gratuitamente, mesmo correndo o risco de me esquecer de alguma delas, quero citá-las: WordSmith, Sphinx, Alceste, IRaMuTeQ, Lexicon[5] e, finalmente, Lexico3. Vou me deter neste último por dois motivos: a) trata-se do sistema que mais domino; b) está disponível gratuitamente para os pesquisadores. Alerto, no entanto, que a melhor ferramenta é aquela com a qual o pesquisador consegue melhores resultados, logo não defendemos o uso de um ou outro sistema, mas que se tenha uma metodologia explicitada, bem construída, como recomendou Pêcheux acima.

Minha intimidade com o Lexico3 foi desenvolvida durante meu doutorado (CONDE, 2008), momento em que usei, além dele, o software Systemic Coder e o gerenciador de banco de dados Access da Microsoft[6], integrando os programas para dar conta de toda problemática do meu objeto de análise. O software *Lexico* está em sua terceira versão desde quando foi criado em 1990 e seu uso envolve um conjunto de procedimentos, não muito complexos, mas bastante trabalhosos que vão desde a preparação do *corpus* até os procedimentos técnicos para extração dos dados, esta por, por sua vez, é uma fase instantânea do trabalho, preparada pelos cálculos automáticos do próprio sistema. Por fim, uma última e mais complexas fase, é a da interpretação dos dados que envolve dois passos: primeiro) extração dos dados a partir das ferramentas disponíveis no sistema; segundo) interpretação dos dados levantados. Vamos, nesta apresentação, apenas demonstrar o primeiro passo.

[5] Ao final das referências bibliográficas listamos os respectivos sítios eletrônicos desses sistemas para futuras consultas.

[6] Minha tese está disponível eletronicamente e recomendo a leitura do capítulo 4, no qual detalho o percurso metodológico com seus erros e acertos.

A preparação do corpus é de extrema relevância porque é o momento que determinará considerável parte do tratamento dos dados. Para um efetivo uso do Lexico3[7] devemos observar que o sistema está pautado sobre um eixo de unidade e um eixo da diferença, pois analisar dados lexicais a partir de textos de qualquer natureza exige do pesquisador um critério que una todos os itens a serem analisados e para tanto várias hipóteses podem ser aventadas enquanto unidade, p. ex. todas as obras de Machado de Assis, todos os artigos de opinião de um dado veículo de comunicação em um determinado período de tempo; todos os discursos presidenciais oficiais em um momento específico do ano, cartas de um personagem histórico, entre outros. No eixo da diferença, se tomarmos as obras de Machado de Assis, podemos identificar aquelas produzidas para diferentes finalidades e/ou gêneros: artigos para jornais, peças teatrais, romances, contos, cartas; sendo cada um desses subgrupos reagrupados em suas massas textuais e etiquetados diferentemente, desse modo a unidade é "Machado de Assis" e a diferença é "gêneros". Se pensarmos nos artigos de opinião de certo veículo, podemos criar os subgrupos cronologicamente (mês, semestre, ano, tema etc.). Poderíamos ainda, observar marcas linguísticas sob a autoria de faixas etárias ou gênero, etc.

No tocante ao processamento do *corpus* o Lexico3 faz sua segmentação da massa textual a partir dos itens lexicais tomando por base o espaço em branco entre as palavras e a pontuação, ou sinais não alfanuméricos de modo que, ao se etiquetar um texto ou grupo de textos, é possível além de fazer o levantamento de termos, observar a sua distribuição. De modo simplificado, a etiquetagem tem também dois eixos: a) critério e, b) item, sendo textualmente colocado antes da parte segmentada da seguinte forma <critério=item>, ou seja, se tomarmos como objetivo analisar textos de opinião poderíamos ter no conjunto de critérios de segmentação do corpus as seguintes configurações: A = {<mês=01>, <semestre=01>, <ano=2011>}; B = {<mês=13>, <semestre=01>, <ano=2012>} ... X = {<mês=yy>, <semestre=0w>,

[7] Veja Conde (2007).

<ano=zzzz>}. Assim, um texto ou segmento do corpus que tivesse em sua etiquetagem a configuração "A" poderia ser comparado a "B" em suas características lexicais e de distribuição desse léxico, permitindo ao pesquisador, se perguntar porque um determinado item, ou conjuntos de itens, ao tratarem do mesmo tema estão em coocorrência entre si ou concorrência de uma parte a outra se considerarmos a diferença entre os anos, entre os meses e semestres. Asssim, sendo critérios e itens moldáveis ao percurso metodológico de análise, as possibilidades se tornam infinitas em termos de comparação.

Observada essa situação preparatória, podemos partir de um brevíssimo estudo de caso, ilustrar as ferramentas disponíveis no sistema Lexico3.

2.1 Estudo de caso

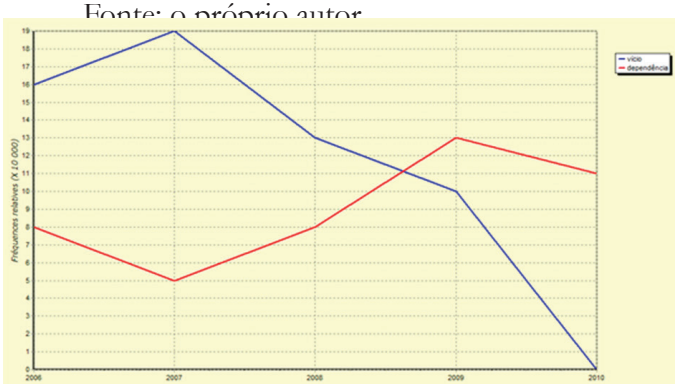
Quero apresentar um estudo de caso a partir de um *corpus* que utilizo com estudantes de graduação e demonstrar, através desse caso, a quantidade de informação necessária para exemplificar como é possível extrair dados. Prefiro exemplificar com esse *corpus,* pois se eu tomasse outro corpus mais complexo teria que utilizar boa parte deste espaço para explicar escolhas metodológicas e a integração com outros sistemas.

Trata-se de um conjunto de 72 textos (aliás uma amostra minúscula!) publicados na seção "Ciência" da Folha de São Paulo, versão on-line, entre janeiro de 2006 e maio de 2010 em que a palavra "cigarro" surgiu, não importasse o tema. Coloquei apenas uma categoria de etiqueta a partir do critério cronológico dos anos de publicação dos textos: <ano=2006> ... <ano=2010>. Falei bastante da etiqueta e aqui, em tempo, vai mais um esclarecimento quanto o sistema Lexico3: ele faz o levantamento quantitativo do léxico e dos sinais e considera, por exemplo, toda a massa de texto compreendido entre <ano=2007> até chegar à próxima etiqueta que pode ser <ano=2007> ou outra qualquer, a ordem não importará.

Por sugestão do auditório vamos então fazer uma comparação entre o uso de termos chaves como "vício" e "dependência", apenas a título de exemplo, ressalvando que esta é apenas uma demonstração de possibilidade e por isso não tem rigor metodológico necessário[8].

A primeira ferramenta a ser empregada demonstra a distribuição de um item lexical ao longo do corpus e ela permite a extração do seguinte gráfico:

IMAGEM 1: distribuição de termos através dos segmentos do corpus - Fonte: o próprio autor



A imagem 1, simplesmente, foi extraída em milésimos de segundos com um clique sobre a ferramenta específica, sem que o pesquisador tivesse que despender tempo em contagens manuais etc. Esse praticidade no entanto não nos isenta de uma difícil pergunta: o que esse gráfico representa?

Basicamente, há, em quase todos os 72 textos, a ocorrência dos termos "dependência" e "vício". Sendo para que para "dependência"

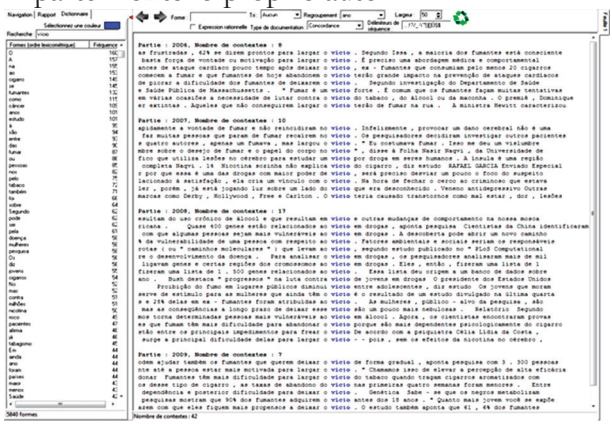
[8] Os termos "vício" e "dependência" têm sentidos e, portanto, efeitos de sentidos distintos, mas para uma melhor análise deveriam ser isolados nos contextos, por isso nossas ressalvas. Para exemplos mais completos e complexos, insisto na consulta do sítio eletrônico da revista *Lexicometrica* <*http://lexicometrica.univ-paris3.fr/*>

foram 29 ocorrências enquanto "vício" teve 42 aparições. No entanto, podemos observar no eixo "y" do plano a frequência dos termos sobre um coeficiente de 10.000 ocorrência, ou seja, proporcionalmente, ao restante de palavras de todo o *corpus* em sua distribuição por ano, demonstra que o termo "vício" declina em uso enquanto o termo "dependência" ascende em relação entre si e todos os temais termos. O que esse tipo de comportamento estatístico pode nos dizer a respeito do sentido e do efeito de sentido desses dois termos que não são sinônimos, mas podem figurar em um "paradigma designacional"[9]? A resposta a essa pergunta depende de uma reflexão profunda, das condições de produção dos textos bem como dos contextos sintáticos e mesmo dos casos que o termo ocupa na estrutura semântica, ou mesmo outro fator diverso a ser levado em consideração. Poderíamos especular que orientações do uso de termos pautadas sobre o "politicamente correto" estejam controlando esses surgimentos ou desaparecimentos. Ou ainda, se justifique essa distribuição por uma questão de modalidade enunciativa e gênero próprios do discurso científico... novamente estamos diante de muitas possibilidades de exploração.

Imaginemos agora que não seja interessante apenas pensar na distribuição quantitativa dos termos chaves, mas é preciso pensar no contexto frasal e suas relações sintagmáticas, semânticas (p. ex. se "vício" e "dependência" são itens que ocupam mormente posições agentivas) e relações anafóricas, posições no parágrafo. Para isso podemos utilizar outros dois recursos do sistema: a) o concordanciador; b) o mapa de seções. O concordanciador colocará os contextos frasais para todas as ocorrências no *corpus*, permitindo que o pesquisador possa isolar contextos de diferentes tamanhos, organizados pela segmentação já discutida anteriormente. Vejamos um exemplo:

[9] Sobre "paradigma designacional", ver Mortireaux (1993)

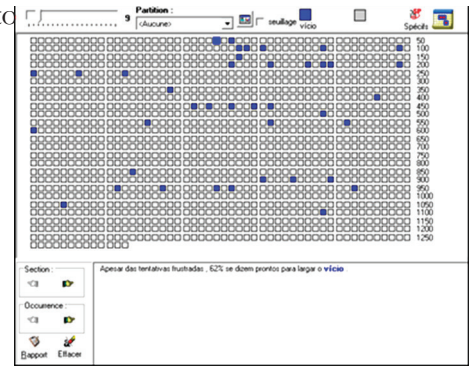
IMAGEM 2: concordanciador aplicado ao termo vício segmentado por parte. Fonte: o próprio autor



Na imagem 2 (novamente fornecida pelo sistema em apenas poucos cliques e milésimos de segundos), temos uma série de informações, mas vamos nos ater apenas à coluna da esquerda e à coluna da direita. Na coluna da esquerda, tem-se todas as formas lexicais mapeadas no *corpus*, não importa seu tamanho: artigos, preposições etc; na coluna da direita temos todos os contextos frasais em que o item "vício" surgiu, no entanto estamos com um contexto pequeno, contando apenas com cinquenta dígitos à esquerda e cinquenta à direita, o que pode ser ampliando em até 999; além disso, vemos que as ocorrências estão organizadas pela etiqueta cronológica determinada no início, de modo que um analista pode ler e analisar os itens conforme sua distribuição em qualquer etiqueta. Todos esses dados podem ser gravados ou copiados em texto, facilitando a forma de exemplificação.

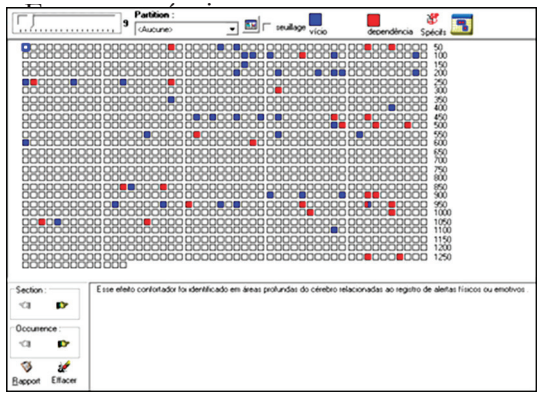
A segunda ferramenta para essa ocasião é o "mapa de seções". Vejamos como ele se comporta:

IMAGEM 3: mapa de seções aplicado ao termo vício. Fonte: o próprio auto



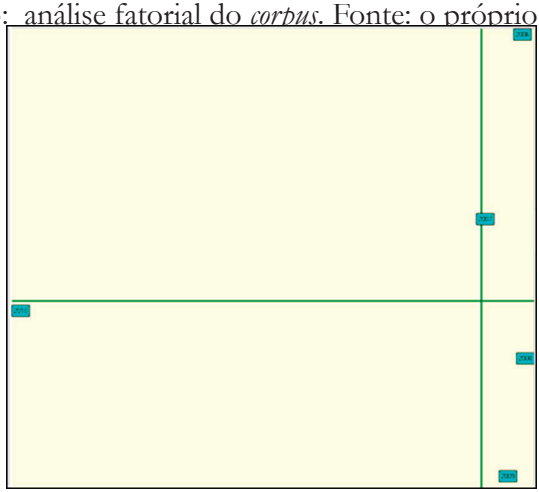
A imagem 3 apresenta a distribuição no corpus todo do uso do termo "vício", cada quadrículo é uma sentença, e os quadrículos azuis, correspondem às ocorrências do termo. Aqui optamos por segmentar o texto em sentenças, mas o pesquisador pode segmentá-lo como quiser e de várias formas simultaneamente: além da pontuação, o pesquisador pode usar a marcação de parágrafo ou outro símbolo não alfanumérico e qualquer modo de segmentação que achar pertinente. Além de essa ferramenta permitir a observação de um item em forma de "mapa", ela ainda permite a verificação da coocorrência de um termo ou grupo de termos. Como podemos observar na imagem 4, na qual, os quadrículos azuis marcam o termo "vício" e os vermelhos marcam "dependência". No entanto, quando se tem os dois termos ocorrendo na mesma sentença, o quadrículo fica dividido entre as duas cores. Também chamo a atenção para a imagem 4, apenas para lembrar que no quadro abaixo do mapa de seções há uma janela que apresenta o segmento selecionado, ou seja, ao clicar sobre qualquer quadrículo, o segmento de texto referente a ele surge; esteja ele marcado ou não. Vale ressaltar que todas essas funcionalidades do sistema podem ser gravadas e recuperadas.

IMAGEM 4: mapa de seções com os termos “vício” e “dependência”.



Uma última ferramenta que gostaríamos de explorar é a “análise fatorial”. Vejamos primeiramente a imagem e em seguida passamos a comentá-la.

IMAGEM 5: análise fatorial do *corpus*. Fonte: o próprio autor



A análise fatorial traz uma representação da distribuição e comparação dos itens lexicais por grupos conforme a etiquetagem. Basicamente, essa ferramenta faz um levantamento das características lexicométricas de cada parte, de modo que o conjunto de termos ou segmentos repetidos sejam colocados sobre um plano cartesiano. Na imagem acima, podemos ler que os textos etiquetados com os anos 2006 e 2007 estão no mesmo quadrante, de modo que 2008 e 2009 também, estando próximos entre si e também observamos que 2010 se distancia dos demais grupos, sendo isolado. Isso, a princípio, pode não dizer muita coisa, mas se tomarmos cada um dos grupos explorando suas semelhanças e diferenças, em termos de ocorrências lexicais, é possível perceber quais são os termos ou segmentos repetidos mais propensos a surgirem e o menos usuais para cada o grupo. Essa ferramenta permitiria um mapeamento preliminar do que classicamente poderia ser chamado de "formação discursiva". Vale lembrar que esse plano cartesiano dever ser interpretado dentro de uma proposta metodológica, porque não se trata apenas de uma imagem, mas de um ponto de partida para outras explorações. Por exemplo, a clicar sobre uma dos segmentos, se juntarmos segmentos por quadrante podemos extrair as listas de termos e segmentos repetidos dentro do *corpus*. Para uma melhor descrição da análise fatorial, recomendamos a leitura de Salem (1982).

Conclusão

Diferentemente do final dos anos de 1960, contamos com muita tecnologia disponível ao alcance de nossas mãos! Os computadores pessoais estão a um custo relativamente baixo e são potentes para lidar com alguns modelos de *corpora* linguísticos e há vários materiais disponíveis em rede. Em relação ao que havia no final do século XX, temos uma infinidade de possibilidades em termos de ferramentas para a composição e manutenção de grandes bancos de dados localmente ou

através da Internet. Do conforto de nosso gabinete ou de nossa casa, podemos acessar bibliotecas, textos, arquivos em diferentes línguas, de diferentes épocas etc. Há sistemas gratuitos patrocinados por entidades de pesquisa disponíveis para a comunidade científica. Se o que Pêcheux e Marindin (1990) afirmaram tinha consistência, mas era impraticável no Brasil dos anos de 1980, por causa das dificuldades tecnológicas, isso já não acontece mais. Se, no entanto, a opção metodológica da AD praticada no Brasil é por uma pesquisa com *corpus* pequeno ou a partir de um arquivo constituído sem necessariamente um rigor metodológico, a recusa a ferramentas como a apresentada aqui se sustenta, porém, se o pesquisador faz uma opção por lidar com *corpus*, faz a opção por lidar com fenômenos linguísticos de qualquer nível ou ordem para poder compreender também fenômenos discursivos, então ferramentas como a apresentada aqui são de grande relevância.

Insisto em dizer que uma análise lexicométrica é apenas uma forma de levantar pistas sobre a materialidade linguística, outras ferramentas e métodos podem ser utilizados. O fato de ser viável ou não, útil ou não para os analistas do discurso depende exclusivamente de sua opção metodológica e também de sua experiência, empiricamente falando. Essa experiência só pode ser avaliada se experimentada, pois simplesmente ignorar não é algo que a ajude a disciplina ou que ajude cientificamente o desenvolvimento da reflexão. Ao participar desta mesa-redonda intitulada *Discurso e "novos" diálogos teórico-metodológicos* não apresentei necessariamente algo "novo", mas talvez algo desconhecido de muitos analistas do discurso em virtude da história da constituição dessa disciplina no Brasil. Longe de mim querer dizer que exista um jeito certo ou errado de fazer AD, mas há no mínimo, um modo linguístico de se fazer AD e sem dúvida é o olhar para sua materialidade. A materialidade pode ser um conceito, mas sua existência é, digamos, material... Nada melhor que qualificar, quantificar e objetivar esse material, mesmo que isso nos leve a nenhuma conclusão, ou a dados negativos o importante é que fizemos

análise e não discurso. Talvez seja uma boa oportunidade para que a língua e sua materialidade saiam da condição etérea, inapreensível dada por uma cultura que exclui a empiria para um fazer teórico-prático mais interessado no fenômeno do que necessariamente em uma bandeira teórica. Por fim, ressalto que o diálogo entre Linguística de Corpus, Processamento de Linguagem Natural e Discurso seria de grande valia para o desenvolvimento e aprimoramento de diversas ferramentas que auxiliariam grandemente o progresso das pesquisas.

Referências

BENZÉCRI, J.-P. et alii. **Pratique de l'Analyse des Données** : linguistique et lexicologie. Paris: Dunod, 1981

CONDE, C. Lexico3 - **Manual resumido de utilização do Lexico3**. Université Paris 3 – La Sorbonne Nouvelle, 2007. Disponível em <http://www.tal.univ-paris3.fr/lexico/lex3-10pas/Lexico3-10premierspas-portugais.pdf>.

_____. **A alternância da referência ao sujeito enunciador e seus efeitos de sentido**. Tese de Doutorado, Universidade Estadual de Londrina. 2008.

FOUCAULT, M. **A ordem do discurso: aula inaugural no Collège de France**. Trad. Laura T. de A. Sampaio, 7. ed. São Paulo: Edições Loyola, 2001.

GUILHAUMOU, J. **L'historien du discours et la lexicométrie** [Étude d'une série chronologique: le « Père Duchesne » d'Hébert (Juillet 1793 - mars 1794)]. In: **Histoire & Mesure**, v. 1 - n°3-4. Varia. pp. 27-46. 1986.

_____. **L'analyse de discours et la texicometrie.** Le Père Duchesne et Le Mouvement Cordelier (1793-1794). In Lexicometrica, n. 0, 1997.

KRIEG-PLANQUE, A. **A noção de "fórmula" em análise do discurso**: quadro teórico e metodológico. (Trad) Luciana Salazar Salgado; Sírio Possenti. São Paulo: Parábola Editorial, 2010.

LEBART, L. e SALEM, A. **Statistique Textuelle**. Dunot, Paris, 1994.

MORTUREAUX, M-F. **Paradigmes désignationnel.** In Semen, n° 8, Paris, 1993. Disponível em <http//semen.revue.org/document4132.html>, acessado em 10 de fev. 2007.

PÊCHEUX, M. E MARANDIN, J-M. **Informatique et Analyse Du Discours**. In L'inquietude Du Discours. MALDIDIER, D. (org.). Paris: Éditions des Cendres, 1990.
_____. in PIOVEZANI, C (Org.); SARGENTINI, V. (Org.) . **Legados de Michel Pêcheux**: inéditos em Análise do discurso. 1. ed. São Paulo: Contexto, 2011.

POSSENTI, S. O dado dado e o dado dado. In PEREIRA DE CASTRO, M. F. **O método e o dado no estudo da linguagem**. Campinas: Ed. Da Unicamp, 1996.

SALEM, A. Analyse factorielle et lexicométrie: synthèse de quelques expériences. In: **Mots**, mars 1982, N°4. p. 147-168.

Anexo I – Lista de Sítios Eletrônicos de divulgação e venda de sistemas

1. WordSmith - http://www.lexically.net/wordsmith/
2. Sphinx - http://www.sphinxbrasil.com/
3. Alceste - http://www.image-zafar.com/
4. IRaMuTeQ - http://www.iramuteq.org/
5. Lexicon – de responsabilidade do Professor José Barbosa Machado – Universidade Trás-os-Montes e Ato Douro (UTAD) - http://www.degois.pt/visualizador/curriculum.jsp?key=2492418931497508

Recebido em 21/11/2014 e Aceito em 15/03/2015.