Publicado em Revista Linguasagem – 16° Edição (www.letras.ufscar.br/linguasagem)

Porque não utilizar ferramentas de informática e estatística em estudos semânticos e discursivos?

Cleber Conde[1]

Introdução

Tradicionalmente há áreas nos estudos da linguagem e em estudos linguísticos que se valem de recursos computacionais e estatísticos sistematicamente. Exemplos disso são a Sociolinguística, a Linguística de Corpus, a Linguística Computacional. Outras áreas, por sua vez, utilizam recursos para análises específicas de *corpus* de modo a suprir uma demanda específica e por fim há áreas e, possivelmente, pesquisadores que não vislumbram aplicações de ferramentas (de informática e estatística) sejam por princípios teóricos ou ainda porque desconhecem a sua eficiência.

Diante de um cenário ainda pouco explorado pelos estudos linguísticos, gostaríamos de lançar um diálogo livre de preconceitos de quaisquer posições, sejam elas contrárias ou favoráveis ao emprego de ferramentas informatizadas e de estatística na análise de *corpus*. Para essa discussão, voltaremos o olhar sobre possíveis aplicações em análises semânticas e discursivas. Este artigo é desafiador para nós porque discutirá possibilidades de tratamento de dados para duas áreas que, no Brasil, não costumam se valer desses expedientes, além disso, o próprio fato de discutirmos aspectos metodológicos nos coloca no centro de boa parte dos problemas de pesquisas, pois acreditamos que muitas teorias e disciplinas já estão consolidadas em suas respectivas áreas de atuação, mas os fenômenos, aos quais se dedicam, apropriando-se desses conhecimentos, nem sempre são comportados e se deixam perscrutar por métodos já experimentados, levando os pesquisadores a lidarem com situações novas.

Para esta discussão iremos traçar um panorama muito geral e, por isso, muito parcial sobre o papel da informática e da estatística textual. Em seguida iremos exemplificar situações: uma aplicação em Semântica e uma aplicação em Análise do Discurso. Inicialmente parecem campos distantes entre si, além disso, são dois campos nos quais os tratamentos informatizados, estatísticos ou automatizados não são bem vistos ou não são tão utilizados, pelo menos, no Brasil.

[1] Docente no Departamento de Letras da Universidade Federal de São Carlos (UFSCar), São Carlos, São Paulo, Brasil: cleberconde@ufscar.br.

Não desejamos convencer nenhum pesquisador a abandonar seus métodos atuais, ou propor uma solução milagrosa que irá resolver impasses em pesquisas. Dar-nos-emos por satisfeitos se o leitor chegar ao final deste artigo e considerá-lo plausível, ou, pelo menos, digno de ser criticado.

Informática e Estatística Textual

Pierre Guiraud faz uma afirmação bastante interessante, em tom de máxima e gracejo: "A Lingüística 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)[2]. Não queremos aqui afirmar que se trata de uma verdade absoluta, incontestável, mas havemos de convir que é uma afirmação bastante provocadora, principalmente, se associarmos a ela um conceito banal de estatística:

> Estatística é o estudo dos modos de obtenção, coleta, organização, processamento e análise de informações relevantes que permitam quantificar, qualificar ou ordenar entes, coleções, fenômenos, populações de modo tal que se possa concluir, deduzir ou predizer propriedades, eventos ou estados futuros. (MILONE, 2004, p.3)

As técnicas quantitativas não são privilégios da Segunda Metade do Século XX, segundo Guiraud (1960, p.5) os estudiosos alexandrinos da Antiguidade já haviam feito o levantamento das *hapax legomena*[3] dos textos homéricos e os massoretas já haviam feito levantamentos sobre a quantidade de palavras dos textos bíblicos hebraicos. Guiraud ainda lembra que publicou em 1954 um levantamento de aproximadamente 2.500 trabalhos organizados em diferentes áreas de interesse dos estudos da linguagem que de algum modo se utilizaram de métodos estatísticos, a maioria sem o aparato tecnológico do qual dispomos atualmente.

Berber Sardinha (2000) lembra que muitas das críticas aos trabalhos sobre *corpora* grandes provinham do tratamento manual de dados, que, em algum momento poderiam ser falseados. Tal situação mudou muito a partir do momento em que entram nas universidades e nos centros de pesquisas os recursos computacionais, de modo que tratamentos manuais passaram a ser substituídos palas máquinas com maior exatidão, rapidez e com *corpora* ainda maiores.

---

[2] La linguistique est la science statistique type ; les statisticiens le savent bien, la plupart des linguistes l'ignorent encore.

[3] Palavras que aparecem apenas uma vez em uma obra.

Muito embora até aqui tenhamos utilizado o conceito de "computacional" aliado à "estatística textual", temos a clareza de que são dois campos possíveis de dissociação. O recurso computacional pode ir desde um editor de texto ou de planilhas até um sistema de inteligência artificial, sintetizadores de voz, passando por sistemas de estatística lexical. Enquanto que as ferramentas de estatísticas podem ser projeções de médias simples até cálculos complexos de curvas e desvios padrões, ou seja, a estatística pode se valer de instrumentos computacionais e vice-versa. Estamos lidando com um universo muito grande de aplicabilidade e por isso, as circunstâncias dos objetivos das pesquisas é que encaminham boa parte dos trabalhos, mas para efeito deste artigo, a estatística e a computação serão amplamente associadas, uma vez que aquela se vale das ferramentas desta, a priori.

Um possível emprego na Semântica

Antes de qualquer discussão sobre o emprego de métodos computacionais e estatísticos em Semântica, devemos pensar de qual semântica estamos falando: Referencial, Verifuncional, do Acontecimento? Inicialmente, o exemplo que trataremos aqui pode ser experimentado em duas possíveis aplicações e em semânticas diferentes: uma referencial, sob a perspectiva da Semântica do Protótipo (Kleiber, 1999) e uma Semântica baseada nos Papéis Temáticos (Cançado 2002, 2005; Chafe, 1979).

Em busca dos protótipos

As semânticas referenciais, como tentativas de compreender as relações mundo-linguagem, esbarram nas relações categoriais, ou seja, que categoria de coisas pode ser chamada de "cão", ou de "fruta"?[4] "Para a fundamental questão *como se categorizam as coisas?*, a resposta clássica, ´aristotélica´, em que se pensa imediatamente é que a categorização se faz à base de propriedades comuns." (Kleiber, 1999, p20)[5]. De acordo com esse objetivo, para se chegar à identidade de um ente, é necessário que se atendam às *condições necessárias e suficientes* (modelo CNS). Por exemplo, para se dizer que um "cão" pertence a sua categoria, são condições

[4] Desde já estamos assumindo que é possível categorizar os seres, diferente de George Lakoff.

[5] "À la question fondamentale *Comment catégorise-t-on?*, la réponse classique "aristotélicienne", à laquelle on pense immédiatement, est que la categorisation se fait sur la base de propriétés communes.

necessárias que ele seja um animal, mamífero, quadrúpede etc., ou seja, impõem-se os atributos e as propriedades comuns de um cão. O problema do modelo CNS está em que "sofre algumas dificuldades para dar conta de 'sentidos múltiplos', ou seja, de palavras que se remetem a vários tipos de referentes possíveis e que apresentam, então, um problema quanto à fixação das CNS." (idem, p. 26).[6]

Como alternativa ao modelo CNS, Kleiber lembra a versão standard da "semântica do protótipo". "O que se chama de *protótipo*? Os testes e experiências descritas nos primeiros trabalhos de E. Rosch introduziram a noção de protótipo como sendo o melhor exemplar ou ainda a melhor instância, os melhores representantes ou instância central de uma categoria." (idem 47-48)[7]. A partir dessa citação, podemos dizer que a Semântica do Protótipo é uma teoria semântica lexical, mas que parte em direção às categorias mais abertas, ou seja, se aplicarmos o modelo de protótipo à noção de pássaro, poderíamos, então, concluir que o pinguim é uma ave, pois ele se aproxima de um modelo de ave (tem plumagem, é ovíparo). No entanto isso não exclui um outro problema a se pensar: se, para João e Pedro, o protótipo de uma ave não é o mesmo, como eles identificam uma ave como tal ou o pinguim como ave? Isso é possível porque, em uma cultura, se tem zonas de saber compartilhadas: "O objetivo da semântica do protótipo é, evidentemente, o de descrever essas zonas de saber prototípico compartilhadas." (Langacker, apud, Kleiber, 1999, p. 49)[8]. Evidentemente estamos lidando aqui com uma visão cognitivista da relação mundo linguagem, no entanto, há de se considerar que, *grosso modo*, "cognescere" do latim é "conhecer", "saber"; assim, atitudes, comportamentos, procedimentos culturais, sociais, ideológicos não deixam de ser um "conhecimento", o fato é que o sentido, para ser compreendido na sua dimensão sócio-discursiva depende um contexto maior, e mais complexo, o que não ignoramos. Kleiber considera que a relação mundo-linguagem também é mediada por aspectos sócio-históricos e condena o objetivismo radical:

> Convém então abandonar a ideia uma existência objetiva da realidade. Nós não temos acesso ao mundo real tal qual ele é. Nós não podemos saber

[6] "(...) Le modèle des CNS éprouve des difficultés à randre compte du 'sens multiple', c´est-à-dire des mots qui renvoient à plusieurs types de référents possibles et qui posent donc um problème quant à la fixation des CNS.

[7] "Qu´appelle-t-on *protoype?* Les tests et expériences décrits dans les premiers travaux d´E. Rosch introduisent la notion de prototype comme étant les meilleur exemplaire ou encore la meilleure instance, les meilleurs représentant ou l´instance centrale d´une catégorie."

[8] Le but de la sémantique du prototype, c'est évidemment de décrire ces zones de savoir prototypique partagé.

> qual é o mundo objetivo nem qual é verdadeiramente a realidade. Como lembra a Gestalt, a realidade é um mundo percebido, uma imagem do mundo, um mundo experimentado, interpretado, construído pela percepção, pela interação e pela cultura. (Kleiber, 1997, p. 12)[9]

Mais tarde, o modelo prototípico sofreu revisões. "Falar de protótipo é simplesmente uma ficção gramatical cômoda; o que é realmente visado são os julgamentos de grau de prototipicalidade." (Rosch, apud, Kleiber, 1999, p. 150)[10]. Assim, diante de uma grande variação de categorias, as significações devem ser compreendidas pela ligação que as categorias mantêm entre si, como se se tratassem de graus de parentesco. O protótipo é um efeito de organização e não o motor organizador da significação. Tanto isso é verificável em situações reais que é comum crianças chamarem um "gato" de "cão", uma vez que o grau de prototipicalidade entre cavalo e cão, para o sujeito em aprendizagem, é bastante próximo[11].

Em se tratando de proximidade de famílias de sentido (idem, p. 157), ou ainda de melhor exemplar, qual seria então o melhor exemplar de "fruta" para um falante de português do brasileiro de determinada região ou classe social? Está algo que trabalhos estatísticos poderiam levantar, com o intuito de responder por determinados protótipos que são construídos sócio-culturalmente. As aplicações de levantamos através de entrevistas, testes com falantes, uma vez quantificados e interpretados podem levar a uma compreensão das construções comuns aos indivíduos, o que abre perspectivas diversas desde trabalhos que vão de aplicações cognitivistas até discursivas. Este é apenas um exemplo de aplicação estatística baseada em uma semântica de uso, na qual compreender eventos referenciais, descritíveis entre os falantes proporciona a tentativa de compreensão de aspectos do uso de uma língua. Neste caso, o empirismo e a pesquisa de campo são úteis desde que as ferramentas de levantamento sejam metodologicamente plausíveis.

---

[9] Il convient donc d'abandonner l'idée d'une existence objective de la réalité. Nous n'avons pas accès au monde tel qu'il est. Nous ne pouvons pas savoir quel est le monde *objectif*, ni quelle est vraiment la réalité. Ce n'est, comme le rappellent les leçons de la Gestalttherie, qu'un monde perçu, une image du monde, un monde expérimenté, interpreété, façonné par notre perception, l'interaction e la culture, que nosu appréhendons.

[10] Parler de *prototype* est simplement une fiction grammaticale commode; ce qui est réelement_visé ce sont lês jugements de degré de prototypicalité.

[11] Parte de nossas considerações sobre a semântica referencial estão também no artigo intitulado "**O sentido de expressões semidescritivas: um estudo semântico-referencial sobre expressões ordinárias" (Conde, 2010).**

Papéis temáticos e ocorrências na linguagem ordinária

Tomemos alguns exemplos retirados de textos que serviram de *corpus* para nossa tese (Conde, 2008)[12]:

(1) Tem sido grande o preconceito contra os negros tanto na sociedade Educacional quanto em outros aspectos.
(2) Esse fato também diminuiria a criminalidade e o envolvimento na drogas , pois só assim os negros sentiriam orgulho de sua cor.
(3) E que devem ser dadas a criação de cotas para o ingresso de negros na universidade?
(4) Os negros têm capacidade de estudar e cursar uma faculdade ou até mesmo ter uma pós, mas antes de tudo isso.
(5) Os negros enfrentam uma série de obstáculos para sobreviverem em meio a uma grande e Cruel população.

Na análise sintática da gramática tradicional, teríamos os seguintes papéis para "negros" em:

- 1, complemento nominal, do nome "preconceito"
- 2, sujeito do verbo "sentir" na oração subordinada
- 3, adjunto adnominal de "ingresso" que é o complemento nominal de criação
- 4, sujeito do verbo "ter"
- 5, sujeito do verbo "enfrentar"

Evidentemente, uma análise gramatical é muito chã, pois não toma os diferentes aspectos ou papéis que o termo poderia ter como nos verbos "sentir", "ter", "enfrentar". O fato é que os papéis semânticos desempenhados pelas expressões que designam as identidades podem ser indícios de como o enunciador enxerga o "outro" de sua identidade.

Pensando nos papéis semânticos, devemos observar que a centralidade das operações fica por conta dos verbos que funcionam como pivôs dos papéis, diferentemente das gramáticas tradicionais que os prevêem partir do léxico por si e não consideram as relações entre os nomes e verbos para a interpretação. Chafe (1979, p. 96) afirma que os verbos têm outro papel:

[12] Tratam-se de textos produzidos por candidatos em um concurso vestibular cujo tema era "as quotas raciais para afrodescentes".

> Minha suposição será a de que o universo conceptual humano total é dicotomizado inicialmente em duas grandes áreas. Uma, a área do verbo, engloba estado (condições, qualidades) e eventos; a outra, a área do nome, engloba "coisas" (tanto objetos físicos como abstrações coisificadas).

Voltando aos exemplos de 1 a 5, vemos diferentes funções para o designativo "negro" que referencia um ser no mundo, mas designar não é o suficiente, é preciso lhe atribuir papéis nesse mundo, em função de ações ou estados ligados a esse ente. Observemos então como Chafe caracteriza as diferenças entre estado, processo e ação. Para tanto, utilizaremos os seus exemplos traduzidos para o português, reconhecendo de antemão que traduções interferem nesse tipo de interpretação, no entanto o que nos interessa é o percurso. Vejamos:

(6) a. The wood is dry. (A madeira está seca.)
b. The rope is tight. (A corda está esticada.)
c. The dish is broken. (A travessa está quebrada.)
d. The elephant is dead. (O elefante está morto.)

(7) a. The wood dried. (A madeira secou.)
b. The rope tightened. (A corda esticou.)
c. The dish boke. (A travessa quebrou.)
d. The elephant died. (O elefante morreu).

(8) a. Michael ran. (Michael correu.)
b. The men laughed. (Os homens riram.)
c. Harriet sang. (Harriet cantou.)
d. The tiger pounced. (O tigre pulou.)

(9) a. Michael dried the wood (Michael secou a madeira.)
b. The men tightened the rope. (Os homens esticaram a corda.)
c. Harriet broke the dish. (Harriet quebrou a travessa.)
d. The tiger killed the elephant. (O tigre matou o elefante.)
(Chafe, 1979, p. 98)

No grupo 6, temos um nome que apresenta um determinado estado. Nesses casos o autor denomina de paciente, ou seja, o elemento que pertence a um estado; já no grupo 7, 8 e 9, os

elementos não pertencem à noção de estado; assim, Chafe caracteriza a diferença entre estado e não-estado, sendo que o não-estado pressupõe a diferença entre "ação" e "processo" que se caracterizam por serem "eventos", em oposição: para um se pergunta "o que é?" e para outro "o que acontece?". Geralmente a oposição funciona, mas isso é falível não servindo de regra, mas de princípio.

No grupo 7 percebemos que o nome mudou de estado ou de condição, por isso se tem um "processo". Nesse caso, ainda é possível dizermos que o nome é paciente, como no estado, já em 8 não se trata de estado, nem mudança de estado, mas de algo que "alguém" faz ou provoca. Já no grupo 9, parece que temos as duas coisas, ação e processo, pois o agente faz algo a um paciente, que no caso tem seu *status* modificado.

O princípio é o de que a compreensão dos verbos nos permite também compreender um pouco do papel dos nomes; logo, podemos perceber três posições básicas: agente, paciente e circunstância, mas que podem melhor ser especificadas se utilizarmos alguns pressupostos de Fillmore[13].

Se os verbos são pivôs dos nomes que os circundam, mas se o status semântico dos verbos varia de acordo com sua natureza, função e valor nas línguas, os nomes, por conseqüência, têm seus papéis afetados. Assim, um nome como "negro" pode ter funções de agente, ou de paciente[14].

Imaginemos um trabalho estatístico que pudesse, sobre um *corpus*, etiquetar verbos e seus argumentos e perceber uma determinada frequência para que se compreenda a "preferência" dos falantes ou enunciadores em atribuir a um léxico nominal a um determinado papel. Nesse caso, um recurso computacional de etiquetagem e categorização de verbos e argumentos nos levaria aquilo que a estatística pressupõe: a descrição e previsão de eventos. Trabalhos como esse já são desenvolvidos, como é o caso do *Berkeley FrameNet*[15]: projeto inicialmente criado por Fillmore para a anotação semi-automizada de casos semânticos e que está disponível na Internet e também o PropBank, ou pouco diferente do FrameNet e elaborado por pesquisadores, dentre eles

[13] Essa incursão pelas categorias semânticas já teve um precedente. Moirand (1988) utilizou-se da metodologia da Gramática de Casos que a ajudou a perceber como os professores de francês, enquanto enunciadores viam as relações de ensino-apredizagem, e daí a possibilidade de se compreender em que tipo de formação discursiva eles se inscreviam e que tipo de posição-sujeito eles ocupavam.

[14] A literatura lingüística sobre a semântica que trata desses casos tem sua origem arraigada nos trabalhos de diferentes pesquisadores dos anos de 1960 e, principalmente, sobre a Gramática Gerativo-Transformacional (GGT), que sofreu diferentes revisões e deu à luz diferentes modelos de análise, como foi com a Gramática de Casos de Fillmore (FILLMORE 1968a, 1968b, 1975); a Teoria da Dependência Conceitual (Conceptual Dependence Theory – CDT) (SCHANK, 1975); a teoria da Semântica Conceitual (Conceptual Semantic Theory – CST) (JACKENDOFF, 1990) e, no Brasil, Franchi (1975).

[15] Disponível no site http://framenet.icsi.berkeley.edu/

Magali Duran[16], que desenvolve um projeto de *Semantic Role Labeling* (SRL - Rotulagem de Papel Semântico) com vistas ao aprendizado de máquinas. Os benefícios de práticas semi-automatizadas ou automatizadas completamente são muito grandes para a compreensão de fenômenos semânticos e sintáticos, inteligência artificial, teorias da informação entro outras aplicações possíveis, chegando as estatísticas de uso que podem ser a entrada para a compreensão de fenômenos semântico-discursivos.

É mister ressaltar que o levantamento, contagem e descrição de objetos em um *corpus*, são fases de outras etapas também importantes e jamais poderiam ser um fim em si mesmas. É necessário desenvolver uma interpretação coerente para os dados, seria muito improdutivo e muito imprudente um levantamento estatístico sem um movimento interpretativo adequado, o que discutiremos mais adiante.

O papel da Estatística em um Trabalho de Análise do Discurso

É interessante lembrar que a Análise do Discurso de Linha Francesa, mais difundida entre os Analistas do Discurso no Brasil, nasce da "paixão" que seu fundador, Michel Pêcheux, nutria pelas máquinas; exemplo disso é sua obra inaugural *Análise Automática do Discurso* (Pêucheux, 1969[1993])[17]. Não iremos entrar em discussões sobre as controvérsias e críticas que a obra de Pêcheux viveu ou vive, mas é mister lembrar que em um de seus últimos trabalhos escritos em colaboração com Jean-Marie Marandin (Pêcheux et Marandin, [1983]1990) ele lamenta as dificuldades e a falta de colaboração entre os pesquisadores nas empreitadas sobre o processamento de línguas naturais e ressalta a importância da apreensão de uma metodologia robusta que possa dar conta de *corpora* volumosos.

> A utilização da informática exige dos analistas do 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, 1990, p. 282)[18]

[16] Para maiores detalhes visitar o site: http://www.nilc.icmc.usp.br/nilc/index.html.

[17] Denise Maldidier lembra essa "paixão" em seu livro L'inquietude du discours, com parte traduzida para o português brasileiro por Eni Orlandi em "*A inquietação do discurso: (Re)ler Michel Pêcheux hoje (2003)*".

[18] La pratique de l'informatique exige des analystes de discours une construction explicite de leurs procédures de description, ce qui est la pierre de touche de leur consitance d'objectes théoriques. Elle

É interessante como a noção de dimensão aparece bastante explícita nessa perspectiva, o que ressalta uma visão de que em algumas circunstâncias "tamanho é documento" sim e a quantidade pode demonstrar e comprovar determinadas hipóteses. Por exemplo, um dos conceitos caros e controversos na Análise do Discurso é o de "Formação Discursiva"[19]: com base em quais aspectos materiais da linguagem podemos dizer que o enunciado "X" pertence à Formação Discursiva "Y"? Ou seria suficiente apenas o gesto interpretativo do analista do discurso?

Estamos diante de uma reflexão sobre a materialidade do discurso e a relação de pertencimento. O conceito de pertencimento é muito caro, jamais deve ser banalizado, ou esquecido, pois é o que indica ao pesquisador os posicionamentos de um sujeito em relação a um outro ou Outro. Seria uma palavra ou uma frase suficiente para demarcar uma Formação Discursiva? Para nós a resposta é não. E nesse caso a quantidade é importante e vamos demonstrar isso. Imaginemos um conjunto de textos de enunciadores de uma dada posição política, sejam esses textos entrevistas, artigos, discursos. Imaginemos ainda que esses sujeitos se posicionam diante de um tema como "relações internacionais" é possível que, sendo de uma mesma orientação política, defendam um determinado posicionamento semelhante entre eles, em oposição a outro grupo político que se coloca contrário. Assim, os enunciados materialmente terão suas semelhanças e diferenças entre os dois grupos e entre os próprios membros do grupo.

É claro que devemos enxergar que os discursos se entrecruzam, estabelecem alianças ou dissociam-se, e ainda poderíamos pensar que em determinados enunciados eclodem dizeres de outros discursos. Aqui, a eclosão se constitui como um acidente e não uma constante. Nos movimentos entre as eclosões, as dissociações, as associações é possível entrever enunciados semelhantes formalmente e significativamente entre os grupos como, por exemplo, relata Maingueneau (2005): qual o caminho para se chegar até as oposições entre os discursos Jansenistas, Humanistas Devotos opostos entre si e estes dois últimos opostos aos discursos dos Protestantes? Evidentemente por oposições materiais entre os enunciados e não se tratam de um

---

permet, en outre, l'appréhension, de corpus varies de grande dimension, ce qui est la pierre de touche de leur validité d'objets descriptifs.

[19] Não nos aprofundaremos na discussão do conceito de Formação Discursiva, para melhor reflexão recomendamos as seguintes leituras: Baronas (2007), Guilhaumou (2004) et Charaudeau e Maingueneau (2004, verbete *formação discursiva,* p. 240)

ou dois enunciados, mas uma massa quantitativamente relevante de enunciados, proporcionalmente ao arquivo constituído.

Defendemos que é possível, do ponto de vista material, realizar mensurações quantitativa e qualitativamente para subsidiar a reflexão do pesquisador. Muitas vezes, enquanto pesquisadores, falamos de *corpus*, método, rigor científico, mas nos esquecemos ou queremos ignorar o fato de que frequência e quantidade de ocorrências podem ser indícios para a interpretação de muitos dados. Por que não pode acontecer na AD?

Os riscos da estatística e das suas interpretações

Vamos partir de um aforismo atribuído ao humorista italiano Pittigrili (Dino Segre): "Estatística: a ciência que diz que se eu comi um frango e tu não comestes nenhum, teremos comido, em média, meio frango cada um." Exageros e caricaturas à parte, sabemos que a Estatística não se presta a esse papel, mas interesses políticos sim, pois dados tomados fora de contexto, sem a devida compreensão podem levar a conclusões desastrosas. Um outro exemplo anedótico foi a pesquisa mal sucedida de um aracnólogo: Conta-se que o pesquisador conseguiu adestrar uma aranha e a cada vez que ele pronunciava uma ordem o aracnídeo se movimentava. O pesquisador então decidiu amputar uma das patas da aranha e depois dar a ordem e ver se ela conseguiria se movimentar. Ele então o fez e a aranha, agora com sete patas, obedeceu à ordem. Intrigado, o pesquisador repetiu a experiência uma segunda vez e o resultado foi o mesmo e assim sucessivamente até restar uma única pata na aranha. Ao comando do aracnólogo, ela, com o maior esforço possível, arrastou-se com sua última pata. Por fim o pesquisador cortou-lhe a derradeira, deu seu comando, mas a aranha não lhe obedeceu o que levou o pesquisador a concluir que os aracnídeos escutam pelas pernas.

O fato de entre sete vezes o experimento do pesquisador ser bem-sucedido não significou que a sua conclusão estava certa, da mesma forma que ele só testou sua hipótese apenas através de um método, além, é claro, da falta de bom senso. O bom senso, tão necessário a qualquer pesquisador, pode parecer uma aposta na subjetividade ou em algo vago, mas na verdade cada situação de pesquisa exigirá do estudioso a dose devida. Na anedota contada acima, o pesquisador era alguém, no mínimo, obtuso indo contra leis lógicas e físicas básicas. Da mesma forma em estudos da linguagem que mobilizem informática e estatística existirão preceitos

lógicos e estatísticos que não necessariamente estejam descritos em seus detalhes nos respectivos manuais, mas figuram na cultura acadêmica, fazendo parte do bom senso, sendo este motivado por, pelo menos, preceitos: a) o da autorreflexão, ou seja, o pesquisador deve rever e questionar seus procedimentos; b) o da resposta provisória, sou seja, uma interpretação não é a última palavra sobre o assunto.

Exemplo bem sucedido (?)

Sabemos que pode parecer presunção de nossa parte, mas gostaríamos de relatar um pouco da nossa experiência que contou com o uso de dois sistemas de levantamento de dados estatísticos em nosso trabalho de doutorado[20] e que pareceu-nos de satisfatório proveito.

Nosso problema inicial dizia respeito à alternância da referência ao sujeito enunciador em um mesmo texto, ou seja, quando se faziam referências à primeira pessoa do singular (1ps), à primeira pessoa do plural (1pp) ou então a nenhuma das duas, utilizando-se a debreagem básica, ou seja disjungir o sujeito da enunciação do seu enunciado: a não-pessoa (np). Assim, seguimos por uma metodologia que procurava perceber a recorrência da alternância, se ela correspondia a uma série identificada de fenômenos e se esses fenômenos tinham relação com o efeito de sentido em textos que versavam indiretamente sobre identidade racial. Como vimos nos exemplos anteriores, eram textos com qualidade formal duvidosa e respondiam a um tema de redação sobre a implementação das cotas raciais para o ingresso de estudantes negros nas universidades.

Depois do amadurecimento de nossas pesquisas, bem como de muitas idas e vindas, desenvolvemos uma metodologia integrada para dar conta de responder a questões que surgiram durante a sua realização: Seria suficiente apenas mapear as alternâncias dentro dos textos, ou seja, apenas dizer quantas vezes um texto passava de 1pp para 1ps e vice-versa? Como os textos falavam de identidade e os sujeitos se posicionavam em oposição ao outro (negro), esse outro aparecia sob o escopo de qual pessoa enunciativa? Dentro do paradigma designacional[21] para negro e para branco, quais seriam as formas mais recorrentes e em quais circunstâncias? Todos

[20] Nossa tese está disponível para baixar no sítio eletrônico da Biblioteca Central da Universidade Estadual de Londrina: http://www.bibliotecadigital.uel.br/

[21] Segundo Mortureux (1993 e 2004, p. 100), o paradigma designacional é a lista de co-referentes, ou seja, de palavras ou de expressões diferentes que fazem referência a um mesmo objeto.

esses dados poderiam ser interpretados para se compreender movimentos discursivos e posições que o sujeito adotaria?

Para responder a essas questões, então integramos o uso do Sistema Lexico3[22] e do Systemic Coder[23], o primeiro para identificar os paradigmas designacionais e o segundo para categorizar os contextos frasais e enunciativos para verificar as séries que poderiam existir no *corpus*. Assim, conseguimos mapear as ocorrências dos designacionais e em que contextos eles mais surgiram. O uso dessas ferramentas integradas nos permitiu uma visão sobre o *corpus* bastante produtiva, de modo a nos auxiliar nas interpretações sobre as alternâncias e, enfim, sobre as formações discursivas presentes nos textos analisados e sem essas aplicações nossa pesquisa talvez não tivesse conquistado resultados interessantes.

Conclusão

Como o título deste artigo sugere, há pelo menos duas leituras possíveis: uma que questiona o não uso dos recursos computacionais e estatísticos e outra que questiona o uso. O título em forma de pergunta é apenas provocativo e talvez seja a indagação que o pesquisador pode fazer a si mesmo no momento de construir seu aparato metodológico. A resposta que o pesquisador terá dependerá da sua construção metodológica e, é claro, pela fundamentação teórica e das próprias características da área de pesquisa. Por exemplo, trabalhos em Gramática Gerativo-Transformacional não teriam, *a priori*, interesse nas ferramentas exemplificadas, tampouco trabalhos em Filosofia da Linguagem, a não ser que esta questionasse o empirismo, entre outros aspectos da abordagem estatística e computacional.

Fora as questões de caráter teórico, existe a decisão do pesquisador, muitas vezes, ela pode estar atrelada as suas "crenças", o seu grau de intimidade com ferramentas computacionais, e mesmo à facilidade de adesão ao novo, ao menos ortodoxo. Há quem "creia" que estatística é uma enganação, ou há quem utilize o microcomputador como uma máquina de escrever

---

[22] Para maiores informações sobre o Lexico3 recomendamos o acesso ao site do http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicowww/, nesse site há um manual resumido de uso traduzido por nós para o português brasileiro.

[23] Para maiores informações sobre o Systemic Coder, recomendamos o acesso ao site www.wagsoft.com.

sofisticada. Discutir crenças e hábitos nem sempre é produtivo, mas questioná-los, refletir sobre eles é bastante saudável à Ciência, o que permite abrir o leque de opções e perceber que existem outros olhares.

O problema enfrentado em qualquer situação nova é a desconfiança que ela sofre. Muito embora estatística lexical e textual não sejam novidades, pelo menos na França; para os pesquisadores brasileiros em estudos da linguagem, é algo, no mínimo, diferente. Se realizarmos uma breve busca em indexadores de artigos científicos na grande área "linguística" utilizando termos relacionados à "computacional" e "estatística", veremos que a quantidade de estudos em língua portuguesa é bem menor que a publicação em outras línguas (inglês e francês, principalmente).

Acreditamos que a escolha ou não de ferramentas de informática e de estatística deve ser feita mediante a opção metodológica fundamentada e que quantificar dados diante de fenômenos não é um "pecado", uma traição aos princípios teóricos, ou um "tecnicismo burocrático". Afinal dados são dados, enquanto informações são *dados* interpretados pelos pesquisadores, por isso, podemos provar que "comemos meio frango" se quisermos enganar; podemos cortar as oito pernas da aranha e acreditar que sua audição se encontra nas patas. No final, a pesquisa e as conclusões podem ser destruídas ou contra-argumentadas pelas falhas no método e pela falta de bom-senso nas análises e nas interpretações. Se um pesquisador tem bons argumentos para sustentar a metodologia de sua pesquisa, da mesma forma o seu crítico deve ter também bons argumentos para refutá-la. Em suma, não podemos dizer nem sim nem não a um método apenas baseando-nos na aventura ou na mera desconfiança.

Referências Bibliográficas

BARONAS, R.L (org). **Análise do discurso: Apontamentos para uma história da noção-conceito de formação discursiva**. São Carlos: Pedro e João Editores, 2007.

CANÇADO, M. Uma Aplicação da Teoria Generalizada dos Papéis Temáticos: Verbos Psicológicos. ***Revista do GEL***. Número Especial: Em Memória de Carlos Franchi..São Paulo: 2002, p. 93-127.

_____. Posições argumentais e propriedades semânticas. **D.E.L.T.A**, São Paulo, v. 25, n. 1., p. 25 - 56, 2005

CHAFE, W. L. **Significado, estrutura e lingüística**. Trad. Francisco da Silva Borba. Rio de Janeiro: Livros Técnicos e Científicos, 1979.

CHARAUDEAU, P.; MAINGUENEAU, D. **Dicionário de Análise do Discurso**. São Paulo: Contexto, 2004.

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

CONDE, Dirceu Cleber. O sentido de expressões semidescritivas: um estudo semântico-referencial sobre expressões ordinárias. **Versão Beta**, anoVIII, abr.jun de 2010.

FILLMORE, Ch.. The case for case. In Bach, E. and Harms, R.T. (orgs.), **Universals in linguistic theory**, Neuw York: Rinehard and Winston, 1968a, p. 1-88.

_____. Lexical Entries for Verbs. *In* **Foundations of Language**, New Iork, 1968, p. 373-393.

Guiraud P. (1960) **Problèmes et méthodes de la statistique linguistique**, P.U.F., Paris.

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

KLEIBER, G. Sens, référence et existence: que faire de l'extra-linguistique? In **Langage**, Paris, n° 127, p. 9-37, 1997

_____.**La sémantique du prototype : catégories et sens lexical**. 2ème Ed. Paris : PUF, 1999.

MAINGUENEAU, D. **Gênese dos discursos**. Trad. Sírio Possenti. São Paulo: Criar, 2005.

MALDIDIER, D. **A inquietação do discurso**: (Re)ler Michel Pêcheux hoje. Trad. Eni Orlandi. Campinas: Pontes, 2003.

MILONE, G. **Estatística: geral e aplicada**. São Paulo: Pioneiro Thomson Learning, 2004.

MOIRAND, S. **Une histoire de discours... une analyse des discours de la revue Le Français dans le monde 1961-1981**. Paris : Hachette, 1988.

PÊCHEUX, M. Análise do Discurso três épocas. In : GADET, F. (org.) **Por uma análise automática do discurso** – uma introdução à obra de Michel Pêcheux. Campinas : Ed. UNICAP, 1993.

_____. E MARANDIN, J-M. Informatique et Analyse Du Discours. In **L'inquietude Du Discours**. MALDIDIER, D. (org.). Paris: Éditions des Cendres, 1990.

SARDINHA, T.B. Linguística de Corpus: histórico e problemática. In **D.E.L.T.A.**, Vol. 16, N.° 2, 2000 (323-367)

SCHANK, R. **Conceptual Information Processing**. Nort-Holland Publishing Company, Nova York, 1975