Simões, Barreiro, Santos, Sousa-Silva & Tagnin (eds.) *Linguística, Informática e Tradução: Mundos que se Cruzam*, Oslo Studies in Language 7(1), 2015. 139–152. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

PESQUISA EM EDUCAÇÃO: PERSPECTIVAS (QUALITATIVAS?) NA EXPLORAÇÃO DE GRANDES CORPORA

MIRIAM LEITE E CLÁUDIA FREITAS

ABSTRACT

Research methods in Education usually rely on qualitative analysis, focusing on samples of individuals or small groups. On the other hand, it is well known that education deals with large scale issues as well: in Brazil, the planning of public policies must take into account the more than 50 million students enrolled in Secondary Education. However, the quantitative approach is viewed with suspicion in Education, leading to very little development of large scale studies. Since these studies can be based on written texts, the dialogue between Education and corpus based approaches becomes highly valuable. In this paper, we briefly present the results of two studies based on corpora specifically designed to foster educational research: (i) a corpus of blogs created and maintained by public schools; (ii) a corpus of teaching materials used in public schools. When discussing the results of these researches, we draw attention to the crucial role played by corpus tools, and to the risks and potentials of the corpus based approach in Education.

Não é preciso ser especialista em Educação para saber que se lida, nesse campo, com questões que se colocam em larga escala: segundo o Censo Escolar da Educação Básica, em 2013, registraram-se 50,04 milhões de matrículas nas redes pública e privada do país. Por outro lado, tampouco é necessária maior *expertise* para se ponderar que o microcosmo da Educação também precisa ser considerado, tanto pela pesquisa acadêmica, quanto pelas políticas públicas. A abstração dos mais de 50 milhões de matrículas se traduz em vida vivida, quando cada uma delas ganha nome e sobrenome e impõe a singularidade da sua localização geográfico-cultural, história familiar, deficiência física ou mental etc. Interessam, portanto, para a pesquisa em Educação, os estudos qualitativos que focalizam tais contingências, mas também investigações e reflexões que operem com dados massivos, que, por certo, são do mesmo modo pertinentes a esse campo.

Entretanto, polêmicas em torno das abordagens quantitativas, que marcaram a pesquisa acadêmica, sobretudo, nas décadas de 1980 e 1990, parecem ainda repercutir na Educação, observando-se pouco desenvolvimento de estudos em larga

escala[1], o que inclui a restrição de pesquisas com grandes volumes de texto organizados em *corpora*.

Diante da crescente multiplicação da produção e da facilitação do acesso a todo tipo de acervo textual, as possibilidades de pesquisa em Educação com esse tipo de focalização são ampliadas e diversificadas, e julgamos que vale problematizar o quadro ainda atual de resistência a pesquisas com acervos empíricos de larga escala. Afinal, os desenvolvimentos tecnológicos que permitiram a disponibilização de um quantitativo informacional inédito na história da humanidade também possibilitaram a criação de ferramentas que viabilizam abordagens inovadoras.

Neste artigo, discutimos o uso de *corpora* na pesquisa em Educação. Com esse propósito, apresentamos o *corpus BlogsSME/RJ*[2] (Leite 2013), para argumentar pela pertinência da utilização de ferramentas de gerenciamento e exploração de *corpus*, como o Corpógrafo (Sarmento et al. 2004), como auxiliares poderosos do pesquisador na exploração do conteúdo de grandes acervos textuais. Trazemos ainda o estudo desenvolvido a partir da análise dos *corpora ApostilasSME/RJCienc* e *ApostilasSME/RJMat* (Romão 2014), como exemplo das especificidades da pesquisa no campo educacional. Em conclusão, assinalamos a necessidade de aproximação entre Educação, Estudos da Linguagem e Linguística Computacional, não apenas para evitar uma apropriação ingênua de tais recursos, como também para criar possibilidades de participar do seu desenvolvimento.

[1] RESISTÊNCIAS E POTENCIALIDADES

Concordando com Gatti (2004)[3], percebemos que é bastante difundido no meio acadêmico da Educação brasileira o entendimento de que, até o período de redemocratização política no país, predominavam as pesquisas quantitativas de viés tecnicista e fundamentação positivista. De fato, em publicação de 1986 que teve grande circulação no campo da Educação – *Pesquisa em Educação: abordagens quali-*

[1] Em recente levantamento realizado a partir da revisão de periódicos publicados em língua portuguesa classificados nas faixas A1 e A2 do sistema Qualis/CAPES (http://qualis.capes.gov.br/) da área da Educação, constatou-se visível crescimento de estudos estatísticos, porém, apenas em viés neotecnicista. Não se trata de análises de *corpora* textuais de grande extensão, mas, sim, de pesquisas em torno dos resultados das avaliações de rendimento escolar em larga escala. Entende-se aqui "neotecnicismo" como uma nomeação genérica para perspectivas educacionais que se caracterizam pelas seguintes marcas: "gestão da vida escolar segundo parâmetros da organização empresarial, com mais profissionais da área da economia e da administração do que educadores atuando no seu planejamento e decisão; centralização das atividades de planejamento pedagógico e alto controle do trabalho docente; concepção de qualidade educacional mensurável em parâmetros estatísticos provenientes de testagem externa à escola, em provas com questões objetivas e padronizadas, em geral restritas às disciplinas de Língua Portuguesa e Matemática; criação de sistema de recompensas para o profissional da educação segundo desempenho dos seus alunos nas avaliações em larga escala, mas também na aprovação na escola; parceria público-privada; atenção às estratégias de marketing na gestão da rede" (Leite (2014), no prelo, nota 12).

[2] O acrônimo SME/RJ refere-se à Secretaria Municipal do Estado do Rio de Janeiro

[3] Esta discussão foi também desenvolvida no artigo *Pesquisa em Educação e cibercultura: questões de metodologia e política*(Leite 2015), para tratar de aspectos políticos que não são aqui focalizados.

tativas (Lüdke & André 2008) – anuncia-se, já na contracapa: "A pesquisa em educação encontra-se atualmente em fase de grande evolução, ampliando seu foco de interesse e métodos para além dos estudos tradicionais do tipo *survey* ou experimental, que constituíram suas mais fortes inclinações durante as últimas três ou quatro décadas."

Entretanto, Gatti (2004) cita estudos que apontam que a pesquisa em Educação era bastante limitada até então e que, nesse universo restrito, apenas 29% operavam com dados quantitativos. Mas o que se observa é que, com ou sem respaldo empírico, difundiu-se, no campo educacional, robusto preconceito contrário aos estudos quantitativos, o que leva a autora a constatar quadro semelhante, passada quase uma década da publicação deste último artigo citado: "tudo o que vem a partir de abordagens 'qualitativas' é bom; o que vem de abordagens 'quantitativas' é mau") (Gatti 2012, pg. 30).

Dificulta-se, assim, a construção de uma crítica mais consistente que permita uma identificação menos apaixonada dos limites e potencialidades da pesquisa com dados massivos. Desse modo, percebe-se a ausência de pesquisadores da Educação quando se desenvolvem tais estudos, que são, com frequência, realizados por profissionais de outras áreas, como especialistas em informática, economistas, administradores de empresas.

Contudo, muitas já foram as vozes da academia que se mobilizaram para matizar tal entendimento e argumentar contrariamente ao reducionismo da antagonização apriorística qualitativo/quantitativo. Brandão (2002), por exemplo, em texto que já conta com mais de dez anos de publicação, argumenta que:

> A incomensurabilidade das práticas sociais não significa, no entanto, que não se possa e deva tentar aproximações quantitativas dos fenômenos. Portanto, os antagonismos quantitativo/qualitativo, assim como micro/macrossocial são improcedentes; informações e dados objetivos, assim como depoimentos e entrevistas em profundidade podem ser produzidos em perspectiva positivista; sem uma conceituação prévia e uma reconstrução a posteriori, nenhum material de pesquisa escapa à superficialidade do mau jornalismo. (Brandão 2002, pg. 28–29).

Ou seja, a associação apriorística entre o trabalho acadêmico com base em dados empíricos de larga escala e abordagens homogeneizadoras e simplistas dos contextos sociais focalizados pela pesquisa em Educação não se sustenta. O reconhecimento da irrepetibilidade do acontecimento social contingente pode nos levar ao estudo do singular, mas também pode se beneficiar do olhar para um quantitativo ampliado de casos singulares.

Santos (2014) faz outra ponderação que julgamos de ainda maior interesse para esta discussão: "a dicotomia entre qualitativo e quantitativo é uma falsa

questão, porque é preciso atribuir qualidades para se poder contar, ou ter pelo menos uma ideia de magnitude". Concordamos e destacamos: além de falsa, é perigosa, pois reafirma a suposta neutralidade dos números. O reconhecimento dos aspectos qualitativos de todo ato de quantificação em pesquisa é fundamental para a desnaturalização das categorias em operacionalização.

Propomos, então, com base nos recursos eletrônicos hoje disponíveis, a busca por uma abordagem qualitativa de dados textuais de larga escala na pesquisa em Educação. Em outras palavras, tentamos desenvolver uma leitura desse tipo de acervo textual por meio das novas tecnologias, em uma perspectiva reconfigurada segundo as especificidades da pesquisa do campo educacional.

De fato, a crescente disponibilização de documentos de interesse para a pesquisa em Educação, sobretudo por meio da internet, impõe urgência na superação desses preconceitos e dificuldades. Documentos públicos, legislação, textos teóricos e literários, registros etnográficos (de observações de campo, fílmicas, televisivas, de redes sociais e outros espaços virtuais de publicação e interação social), transcrições de entrevistas e matérias jornalísticas são apenas alguns exemplos dos textos que podem interessar ao pesquisador da Educação. Até o momento, predomina a abordagem manual dessa empiria, o que obviamente limita o escopo dos estudos desenvolvidos.

Para argumentar pela pertinência do acesso à integralidade dos *corpora* cuja extensão compromete a possibilidade do seu processamento por meio da leitura convencional, apresentamos, a seguir, o *corpus BlogsSME/RJ*, para compararmos os estudos desenvolvidos a partir de leitura amostral, com sua posterior abordagem digital, que possibilitou acesso à totalidade do *corpus*.

[2] LEITURAS DIGITAIS

O estudo que deu origem ao *corpus BlogsSME/RJ* foi desenvolvido no contexto da pesquisa *Diferença e desigualdade na educação escolar do jovem adolescente: desconstruções*, em que se indagava acerca dos sentidos de juventude e adolescência afirmados em contextos virtuais de publicização de atividades escolares dos anos finais do ensino fundamental da rede pública municipal do Rio de Janeiro. Tendo-se constatado, em estudo anterior, o estímulo, por parte da Secretaria Municipal de Educação do Rio de Janeiro/SME-RJ, à utilização dos recursos digitais de comunicação contemporâneos, supusemos que os blogs das escolas municipais cariocas que atendem aos anos finais do ensino fundamental poderiam conter registros relevantes relativamente às identificações desses estudantes. Por meio do portal RioEduca[4], organizado pela SME-RJ e responsável pela disponibilização dos blogs da sua rede de ensino, foram selecionados aqueles relativos aos anos finais do ensino fundamental, que atendem à faixa etária priorizada em nossos estudos.

[4] http://www.rioeduca.net

Chegou-se, então, a um conjunto de 160 blogs, ativos entre janeiro de 2009 – início da gestão da SME/RJ que promoveu a criação e desenvolvimento desses blogs – e novembro de 2013, quando se realizou a pesquisa. Destes, 100 eram blogs de escolas, 14, de projetos específicos, 30, de professores, 01, da 5ª Coordenadoria Regional de Educação/CRE[5].

Devido ao grande volume de documentos compilados, a leitura inicial desse material teve de se restringir ao quantitativo possível na abordagem convencional: foram selecionados 20 blogs, incluindo blogs de escolas, de professores e de projetos específicos, de todas as 11 coordenadorias regionais de Educação. Foi feita a leitura extensiva de todas as postagens, comentários e da seção "Quem somos", buscando destacar registros de interesse para a pesquisa.

Não foram localizados registros significativos de atenção a *juventude* e *adolescência*, focos da pesquisa. Predominava a postagem de fotos que pouco contavam sobre a identificação atribuída aos estudantes adolescentes nos contextos retratados. Havia muitas imagens de atividades esportivas, formaturas, exposições de trabalhos ou mesmo dos "Aniversariantes do mês", como nas reproduções que se seguem (figuras 1 e 2). As imagens eram postadas com poucas informações que esclarecessem sobre seu desenvolvimento e propósitos, e quase sempre não se seguiam comentários. Entendeu-se então que o estudo confirmava as conclusões de outras pesquisas: os anos finais do ensino fundamental geralmente não têm sido atendidos em suas especificidades[6] – entre a escolarização da infância nos anos iniciais do ensino fundamental e a profissionalização e/ou preparação para o ingresso na universidade, que têm lugar no ensino médio, a educação escolar do estudante adolescente no ensino fundamental parecia não estar recebendo a devida atenção em políticas públicas ou nas práticas escolares cotidianas. Mas questionou-se também a pertinência dessa empiria – blogs disponibilizados no portal RioEduca –, que pareceu ser de interesse bastante restrito.

Neste momento, o contato com as ferramentas para gerenciamento e manipulação de *corpora* eletrônicos apareceram como uma alternativa a ser investigada. Procedeu-se então à compilação de todo o conteúdo desses blogs, de modo a serem processados por programas como o Corpógrafo (Sarmento et al. 2004). O *corpus* assim construído contém todas as postagens e comentários dos 160 blogs, no período de janeiro de 2009 a novembro de 2013, além do conteúdo da seção "Quem somos", totalizando mais de 4 milhões de palavras[7].

De início, entre os vários programas disponíveis, optou-se pela utilização do Corpógrafo, por razões de ordem prática, mas que também tinham conteúdo político: a opção por um programa gratuito, de uso público e em língua portuguesa, não apenas facilitava o trabalho, como implicava posicionamento político de re-

[5] Subdivisão administrativo regional da SME/RJ

[6] Como por exemplo em Davis et al. (2013).

[7] A documentação completa e o *corpus* estão em `http://www.ddeej.com`

FIGURA 1: Reprodução de blog utilizado na pesquisa.

FIGURA 2: Reprodução de blog utilizado na pesquisa.

levo, na medida em que fortalecia iniciativa em prol do acesso livre a recursos digitais, dados e metadados, e de resistência à hegemonia da língua inglesa nos ambientes virtuais. Contudo, traremos para este artigo os dados gerados pelo programa AntConc (Anthony 2012)(gratuito, de propriedade privada, em língua inglesa), posto que, em 2014, o acesso online ao Corpógrafo esteve irregular.

Na compilação dos conteúdos dos blogs, as fotos foram substituídas pela palavra "foto" e os vídeos, pela palavra "vídeo". Ao ordenarmos as palavras pela sua frequência (figura 3), a palavra "foto" despontou como das mais recorrentes. Pareciam se confirmar, desse modo, as conclusões a que se havia chegado com a leitura exaustiva da amostra dos blogs em estudo.

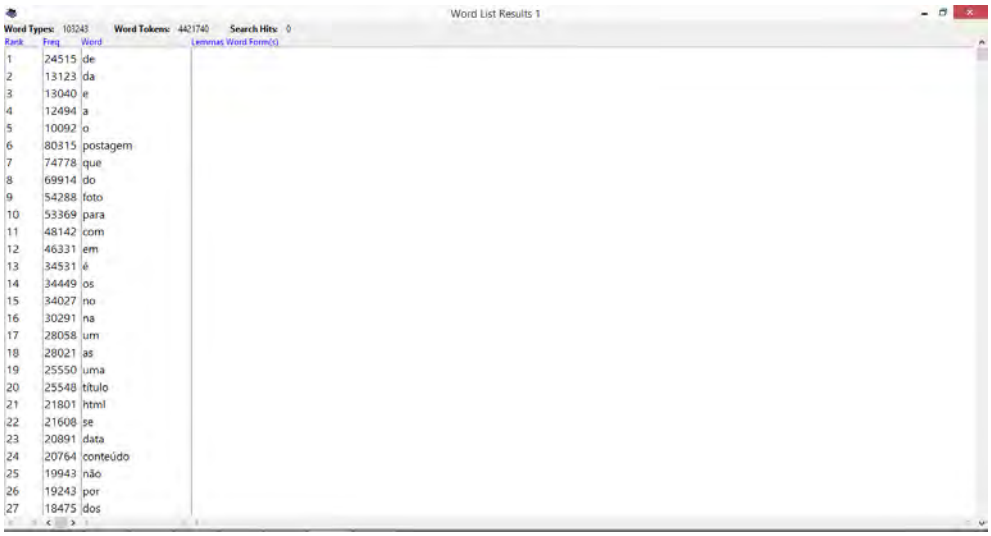

FIGURA 3: Ordenação de palavras por frequência no *corpus BlogsSME/RJ*

Entretanto, apesar de não terem sido localizadas em ocorrências significativas quando da leitura inicial, palavras de óbvio interesse para a pesquisa apareceram na ordenação da listagem de palavras por frequência. Assim, foi possível acessar 2200 repetições das palavras *jovens/jovem*, 1940 para *adolescentes/adolescente*, 270 para *juventude/juventudes*, 239 para *adolescência*, o que evidenciou mais do que a pertinência dessa empiria: demonstrou uma riqueza inacessível sem o auxílio de recursos das tecnologias digitais.

Não caberia aqui trazer todas as reflexões oportunizadas pela problematização dessas palavras e seus contextos de enunciação. Analisando as linhas de concordância, pudemos concluir que, no Rio de Janeiro, não se confirmava a tendência mais geral de não reconhecimento de especificidades dos anos finais do ensino fundamental. Pelo contrário, havia clara atenção direcionada a essa faixa etária, com conteúdo político que valia problematizar. Por exemplo, constatou-se que, sob a expressão *protagonismo juvenil* e afins, desenvolviam-se atividades diversas que trabalhavam pela formação de uma juventude cuja inserção social é pautada por uma perspectiva individualista e neoliberal. A partir deste achado, invisível na leitura parcial dos blogs, foi concebido novo projeto de pesquisa[8], orientado à discussão de tais opções de formação escolar pública dos jovens adolescentes.

Muitas vezes, no entanto, palavras com uma única ocorrência podem ter valor para a pesquisa. Nesse caso, o acesso digital à integralidade dos textos torna-se

[8] Pesquisa *O grêmio e outros espaços-tempos de diálogo político na escola: possibilidades contemporâneas*, contemplada com financiamento pelo Edital Apoio à Melhoria do Ensino em Escolas da Rede Pública Sediadas no Estado do Rio de Janeiro – 2014.

ainda mais produtivo, como se argumenta na próxima seção, a partir das conclusões do estudo dos *corpora ApostilasSME/RJCienc* e *ApostilasSME/RJMat* (Romão 2014).

[3] QUANDO A AUSÊNCIA CRIA SENTIDO

Também no contexto da pesquisa *Diferença e desigualdade na educação escolar do jovem adolescente: desconstruções*, (Romão 2014) desenvolveram estudo sobre as repetições e deslocamentos em torno dos sentidos do feminino presentes nas apostilas distribuídas pela SME/RJ para os anos finais do ensino fundamental – 7º, 8º e 9º ano – sob o nome *Cadernos Pedagógicos*. Trata-se de material didático amplamente utilizado na rede pública carioca, posto que seu conteúdo pauta as avaliações externas municipais e nacionais, condicionando rankings e respectivas recompensas materiais e subjetivas.

As apostilas dos 4 bimestres letivos de 2013 de todas as disciplinas ficaram disponíveis[9] nesse período e foram organizadas, por disciplina, de modo a constituir *corpora* com a íntegra dos conteúdos dos Cadernos Pedagógicos. Embora não tão extensos quanto em geral se apresentam os *corpora* dos estudos linguísticos, sua exploração por meio das ferramentas específicas evidenciou mais uma vez a potencialidade desse tipo de abordagem.

Entendia-se, com base em proposições da teórica feminista Judith Butler (Butler 2003, 1997), que a identidade de gênero se constrói performativamente, isto é, não se constitui em decorrência de marcas biológicas, mas, sim, pela constante e difusa repetição do que socialmente se concebe como característico de cada gênero. Interessavam, portanto, não apenas as passagens das apostilas em que a temática do gênero era explicitamente tratada, como também e sobretudo, aquelas onde, de forma naturalizada, se reafirmavam e/ou se deslocavam os modos do feminino na nossa sociedade. Desse modo, a exploração do material didático em toda a sua extensão parecia especialmente importante. Destacamos, a seguir, duas das conclusões oportunizadas por essa abordagem, que entendemos exemplificar potencialidades de uma outra maneira de leitura de grandes acervos textuais na pesquisa do campo educacional.

O primeiro destaque diz respeito ao *corpus* de Ciências (*ApostilasSME/RJCienc*). Na leitura exploratória dessas apostilas, chamou nossa atenção que as palavras *brasileira/brasileiras* tinham quase a mesma frequência de ocorrência que a sua variação no masculino. No entanto, quando acessamos os contextos de enunciação dessas palavras, por meio da leitura das linhas de concordância, identificamos flagrante desigualdade no valor político-cultural dessas referências.

Enquanto a expressão no feminino qualificava a população residente no país, espécies nativas e práticas culinárias, sua versão no masculino lembrava, na maior

[9] http://www.rio.rj.gov.br/web/sme/material-pedagogico

parte dos casos (7 ocorrências) feitos de cientistas. Quanto a cientistas brasileiras, houve uma única referência. A seguir listamos alguns dos contextos:

- A culinária brasileira é bem original e diversificada. (Caderno Pedagógico de Ciências, 8º ano, 2º bimestre, 2013)
- Faça uma pesquisa sobre a variedade de aves brasileiras e seus cantos distintos. (Caderno Pedagógico de Ciências, 9º ano, 4º bimestre, 2013)
- Foi a primeira brasileira a fazer o concurso para ser naturalista do Jardim Botânico e foi aprovada em 2º lugar. (Caderno Pedagógico de Ciências, 7º ano, 3º bimestre, 2013)
- MICHAEL NICOLELIS (1961), médico, esse brasileiro é considerado um dos 20 maiores cientistas mundiais da atualidade. (Caderno Pedagógico de Ciências, 8º ano, 1º bimestre, 2013)
- O brasileiro Santos Dumont realizou o primeiro voo com o 14 BIS. (Caderno Pedagógico de Ciências, 9º ano, 1º bimestre, 2013)
- A doença de Chagas afeta órgãos como o coração e os intestinos e foi descoberta pelo médico brasileiro em abril de 1909. (Caderno Pedagógico de Ciências, 7º ano, 2º bimestre, 2013)

A leitura convencional das apostilas, no entanto, talvez ocultasse esta e outras que consideramos importantes reiterações da invisibilização da mulher na Ciência. Observe-se que, na apostila do 8º ano, encontra-se uma seção destinada a problematizar as relações de gênero, ali anunciadas como socialmente construídas. Julgamos possível que a explicitação da problemática do gênero se destacasse mais do que suas menções fora dos holofotes textuais ao longo da íntegra do material didático.

Do mesmo modo, no *corpus* que reuniu as apostilas de Matemática, ocorrências singulares nos deram importantes pistas para se compreender o papel da educação escolar na perpetuação do sexismo na nossa sociedade. Consultando a lista de palavras do *corpus* ordenadas por frequência, encontramos, nas últimas posições, nomes próprios femininos e masculinos, e buscamos seu contexto de enunciação. Descobrimos que, em geral, se tratava de personagens dos tradicionais problemas de Matemática, que reiteravam estereótipos masculinos e femininos:

- Miriam quer fazer um bolo grande, aumentando, proporcionalmente, a quantidade de ingredientes. (Cadernos Pedagógicos de Matemática, 7º ano, 4º bimestre, 2013)
- Em uma partida de videogame, Aurélio conseguiu 160 pontos em três rodadas. (Cadernos Pedagógicos de Matemática, 8º ano, 2º bimestre, 2013)

Concluímos, nesse estudo, pela importância da atenção às nomeações cotidianas do gênero, para além da sua discussão explícita e focalizada. Seu poder de naturalização é considerável, na medida em que, ao trazer tais afirmações de modo periférico ao tema central do texto, encontra um interlocutor desprevenido, que tende a ponderar menos os enunciados a que se expõe, posto que não se colocam na direção primeira da sua atenção. Mas concluímos também que, para acessar essas repetições do dia a dia, é importante assegurar uma leitura mais abrangente e sistemática do que o possível manualmente. De fato, a leitura de textos de interesse para a pesquisa em Educação pode se beneficiar de abordagens que também levam em conta aspectos quantitativos do conteúdo, apropriando-se das ferramentas e utilizando-as de modo a enriquecer as formas tradicionais de análise.

[4] APROPRIAÇÕES: RISCOS, LIMITES E PERSPECTIVAS

Para além do tratamento prioritariamente quantitativo oferecido pelo Corpógrafo e programas similares, estudos que fazem uso de grandes *corpora* em áreas que não tematizam diretamente a linguagem começam a surgir, como indica o crescimento do campo das Humanidades Digitais. Com respeito ao diálogo com o campo educacional, especificamente, finalizamos com algumas considerações acerca do que denominamos como *riscos*, *limites* e *perspectivas*.

Sobre os riscos da pesquisa com *corpus* em Educação, destacamos que a prática da utilização de *corpora* eletrônicos não deve ser incorporada ingenuamente, sem levar em conta discussões a que pode estar associada no campo da linguagem, dado o risco de fragilizá-la por incoerência teórica.

Como exemplo, podemos citar o alinhamento a abordagens chamadas *corpus-driven* ou a abordagens *corpus-based*, que dizem respeito sobretudo ao papel atribuído ao *corpus* em sua relação com a teoria.

Vale lembrar que, na Linguística, boa parte dos estudos sobre a linguagem se sustentava em dados provenientes de pelo menos uma das seguintes fontes: intuição do falante; testes de aceitabilidade/usabilidade; entrevistas com informantes. Assim, o uso massivo de grandes *corpora* eletrônicos é saudado como recurso capaz de revolucionar o estudo e descrição da língua, quer propondo novos modelos, quer validando ou refinando modelos já existentes (Sampson 2001; de Beaugrande 2002).

Em geral, quando se usa o termo *corpus-driven* (guiado ou conduzido por *corpus*), assume-se o *corpus* como espaço que viabiliza uma observação neutra dos fatos da língua, que, por sua vez, irá promover a criação de hipóteses. A língua é vista como um fenômeno probabilístico (e daí a relevância de *corpora* grandes), cabendo à exploração com *corpus*, em última análise, a substituição ou revisão de teorias de linguagem, porque erguidas sobre bases inadequadas, ou estabelecimento de novas dimensões de descrição.

Na visão chamada *corpus-based*, o *corpus* é o espaço para validação, refutação ou refinamento de hipóteses prévias, de perguntas previamente formuladas.

A essa diferente maneira de perceber o *corpus*, podem corresponder, também, diferentes posicionamentos com relação às possibilidades do fazer científico.

Perspectivas *corpus-driven* costumam estar vinculadas à aplicação de testes estatísticos, sobretudo quando se trata da descrição/observação de fenômenos mais vinculados ao sentido das palavras ou expressões. Tais testes estatísticos seriam capazes de extrair resultados mais "objetivos" - porque obtidos sem a interferência humana e sem as limitações da intuição. A responsabilidade de responder às questões de pesquisa é transferida para o *corpus*; o pesquisador apenas informa o que o *corpus* "revela", o que veio à tona por meio da exploração automática.[10]

Outra característica comum a essa abordagem é aposta na impossibilidade de atribuição de sentidos das palavras fora de seus contextos de uso, estando esse contexto refletido no *corpus* – daí o destaque para estratégias vinculadas à procura por padrões de uso e padrões de co-ocorrência e à extração de n-gramas.

Abordagens *corpus-driven* podem se associar, ainda que involuntariamente, aos seguintes pressupostos: (i) crença na objetividade e na neutralidade do pesquisador, que não atua sobre os dados, apenas relata resultados de "experimentos"; (ii) crença na possibilidade de um sentido estável, intrínseco às palavras e expressões, que está *no* texto (ou contexto), ou seja, *no corpus* – e o *corpus* é "confiável". É sobre o *corpus* que o pesquisador atuará, fazendo uso das ferramentas adequadas, tendo em vista revelar/extrair sentidos.

Tais pressupostos são respaldados pelo que a reflexão desconstrutora chama de tradição logocêntrica, caracterizada por separações claras e objetivas entre pares dicotômicos e hierárquicos como sujeito e objeto, leitor e texto, literal e metafórico, significado imanente e significado acidental, ironia e não-ironia literariedade e não-literariedade, os quais nenhuma teoria da linguagem conseguiu, até hoje, distinguir de maneira incontroversa (Arrojo 1992).

Quando constatamos que, no diálogo com o campo educacional, a reflexão desconstrutora tem comparecido com alguma frequência (e com mais frequência do que nos estudos da linguagem, como nos lembram Arrojo (1992) e Martins (1999)), as considerações sobre o uso de *corpus* e sua relação com perspectivas de linguagem e de conhecimento não podem ser ignoradas, quando se valoriza a coerência e a consistência da fundamentação teórica da atividade de investigação científica. Sabemos que o pesquisador não é neutro, tampouco o são as ferramentas.

Sobre os *limites*, observamos que programas como o Corpógrafo são de grande valia nas primeiras aproximações de *corpora* mais extensos, sendo capazes de indicar pistas e caminhos que serão explorados por meio da leitura convencional.

[10] No entanto, como obeservamos por Sampson (2001), nem sempre a ênfase na objetividade dos dados obtidos com *corpus* está associada a uma perspectiva *corpus-driven*, e nem esta última está, necessariamente, vinculada à aplicação de testes estatísticos.

No estudo do *corpus Blogs SME/RJ*, foram obtidas 2200 linhas de concordância para as palavras *jovem/jovens*, e 1940 para *adolescente/adolescentes*, implicando tempo significativo para o acesso, caso a caso, dessas inscrições – constatação que nos leva ao que propomos como *perspectivas*.

A aproximação com os estudos linguísticos com *corpus* leva aos trabalhos com *corpora* anotados, do qual o serviço do AC/DC (Costa et al. 2009) é exemplar: voltando à pesquisa sobre as questões de gênero, a leitura das linhas de concordância para verificar os contextos de brasileiro(s) em oposição a brasileira(s) ganharia novos contornos com a observação da distribuição dos substantivos modificados por cada um dos itens mencionados. Do mesmo modo, para os personagens dos problemas de Matemática, seria vantajoso poder buscar diretamente por nomes próprios que se referem a pessoas (e não a lugares ou instituições, por exemplo).

Assim, para além do tratamento prioritariamente quantitativo oferecido pelo Corpógrafo e programas similares, entendemos que a anotação linguística de textos, a partir de questões específicas da pesquisa em Educação, pode viabilizar o trabalho de discussão sistemática de grandes volumes textuais (Freitas 2014). Trata-se de projeto multidisciplinar, que depende da mútua aproximação entre os Estudos da Linguagem, a Linguística Computacional e a Educação, o que certamente não se efetiva em curto prazo. Acena, no entanto, com a possibilidade de ganho que parece valer os custos que coloca: a possibilidade da abordagem qualitativa de *corpora* de larga escala.

Como nossas últimas palavras, lembramos que inquietação e curiosidade fazem parte do perfil do pesquisador, e a apropriação do Corpógrafo que apresentamos aqui ilustra esse aspecto: idealizado por Belinda Maia, foi o primeiro programa com que tivemos contato para verificar as possibilidades de uma abordagem alternativa dos textos da pesquisa em Educação, mesmo tendo sido criado com o objetivo de auxiliar a tradução e a gestão de terminologias.

Não custa portanto imaginar um cenário ideal para a pesquisa com grandes *corpora* que conjugasse as ideias inicialmente concretizadas no Corpógrafo (compilação e gerenciamento dos próprios *corpora*, de interesse do pesquisador/a) e as facilidades do AC/DC – anotação morfossintática e semântica, sistema de busca e serviços que a ele vem se associando (Santos 2014). Vale lembrar que se tais serviços vêm sendo desenvolvidos no contexto dos estudos da língua, não é improvável que outros usos surjam daí, repetindo o próprio uso de *corpora* e do Corpógrafo, situação favorecida quando se tem recursos de qualidade públicos, abertos e disponíveis - novas apropriações, novos usos.

REFERÊNCIAS

Anthony, Laurence. 2012. AntConc (version 3.3.5). http://www.antlab.sci.waseda.ac.jp.

Arrojo, Rosemary (ed.). 1992. *O signo desconstruído*. Pontes.

de Beaugrande, Robert. 2002. Descriptive linguistics at the millennium: corpus data as authentic language. *Journal of Language and Linguistics* 1(2). 91–131.

Brandão, Zaia. 2002. *Pesquisa em educação: conversas com pós-graduandos* Coleção Teologia e ciências humanas. Editora PUC-Rio.

Butler, Judith. 1997. *Excitable speech. A politics of the performative*. Routledge.

Butler, Judith. 2003. *Problemas de gênero: feminismo e subversão da identidade*. Editora Civilização Brasileira. Tradução de Renato Aguiar.

Costa, Luís, Diana Santos & Paulo Alexandre Rocha. 2009. Estudando o português tal como é usado: o serviço AC/DC. Em *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*, s/pp.

Davis, Claudia Leme Ferreira, Gisela Lobo Baptista Pereira Tartuce, Patrícia C. Albieri de Almeida & Ana Paula Ferreira da Silva. 2013. Os esquecidos anos finais do ensino fundamental: políticas públicas e a percepção de seus atores. Em *Anais da 36a Reunião Anual da ANPEd*, .

Freitas, Cláudia. 2014. Corpus, Linguística Computacional e as Humanidades Digitais. Em Miriam Leite & Carmen Gabriel (eds.), *Linguagem, Discurso, Pesquisa e Educação*, 22–51. DP et Alii.

Gatti, Bernardete. 2004. Estudos quantitativos em educação. *Educação e Pesquisa* 30(1). 11–30.

Gatti, Bernardete. 2012. A construção metodológica da pesquisa em educação: desafios. *Revista Brasileira de Política e Administração da Educação* 28(1). 13–34.

Leite, Miriam. 2013. Blogs SME/RJ. http://www.ddeej.com.

Leite, Miriam. 2014. Adolescência e juventude em desconstrução: textos e contextos na educação escolar. Em Miriam Leite & Carmen Gabriel (eds.), *Linguagem, Discurso, Pesquisa e Educação*, 281–307. DP et Alii.

Leite, Miriam. 2015. Pesquisa em educação e cibercultura: questões de metodologia e política. Em Edméa Oliveira & Maria Luiza Oswald (eds.), *Educação, cibercultura e redes sociais em tempos de mobilidade*, no prelo.

Lüdke, Menga & Marli André. 2008. *Pesquisa em educação: abordagens qualitativas* Temas básicos de educação e ensino. EPU.

Martins, Helena. 1999. *Metáfora e polissemia no estudo das línguas do mundo: uma apresentação não representacionista*: Universidade Federal do Rio de Janeiro. Tese de Doutoramento.

Romão, Carla de Oliveira. 2014. *Identificações do feminino em materiais didáticos contemporêneos*: Universidade do Estado do Rio de Janeiro. Tese de Mestrado.

Sampson, Geoffrey. 2001. *Empirical linguistics*. Continuum.

Santos, Diana. 2014. Podemos contar com as contas? Em Sandra Aluísio & Stella Tagnin (eds.), *New language technologies and linguistic research: a two-way road*, 194–213. Cambridge Scholars Publishing.

Sarmento, Luís, Belinda Maia & Diana Santos. 2004. The Corpógrafo - a Web-based environment for corpora research. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, 449–452.

CONTACTOS

Miriam Soares Leite
Universidade do Estado do Rio de Janeiro
miriamsleite@yahoo.com.br

Cláudia Freitas
PUC-Rio
claudiafreitas@puc-rio.br