Edital MCT/CNPq N º 14/2009 – Universal

1- IDENTIFICAÇÃO DA PROPOSTA
TÍTULO:
PADRÕES DO PORTUGUÊS POPULAR ESCRITO: O VOCABULÁRIO DO JORNAL *DIÁRIO GAÚCHO*. FASE 1
Abreviatura**: PorPopular Fase 1**

ÁREAS IMPLICADAS:
LINGÜÍSTICA DE *CORPUS*, LEXICOLOGIA, ESTUDOS DA LÍNGUA PORTUGUESA, ESTUDOS DO TEXTO E DO DISCURSO, LINGÜÍSTICA APLICADA, PROCESSAMENTO DA LINGUAGEM NATURAL.

PROPONENTE/RESPONSÁVEL:
Profa. Dra. Maria José Bocorny Finatto (UFRGS)

Duração prevista: 24 meses.

Equipe de pesquisa:
UFRGS/Inst. de Letras
Prof. Dra. Maria José Bocorny Finatto (coordenadora)
Prof. Dr. Valdir do Nascimento Flores (colaborador)
Profa. Dra. Carmem Luci da Costa Silva (colaborador)
Kleber Valenti Schenk (graduado em Letras, voluntário de pesquisa)
Bruna Rodrigues da Silva (estudante de graduação de Letras)
Daniel da Costa Silva (mestrando do PPG-Letras da UFRGS)
Vera Maria Araujo Pigozzi de Araujo (doutoranda do PPG-Letras da UFRGS)

UFRGS/Instituto de Informática
Profa. Dra. Aline Villavicencio (colaborador)
Faculdade de Informática - FACIN-PUC-RS
Profa. Dra. Renata Vieira (colaborador)
NILC-USP Núcleo Interinstitucional de Lingüística Computacional
Profa. Dra.Sandra Aluísio (colaborador)
UNISINOS
Profa. Dra. Marlene Teixeira (colaborador)
UERGS – Universidade do Estado do Rio Grande do Sul
Profa. Dra. Magali Endruweit

CARACTERIZAÇÃO SINTÉTICA DA PESQUISA A SER EMPREENDIDA

Quadro Geral da proposta

Esta pesquisa pretende obter uma descrição e estudos de padrões do vocabulário exibido por textos de jornais diários populares brasileiros voltados para públicos de menor poder aquisitivo. A descrição e os estudos serão feitos à luz de referenciais teórico-metodológicos da Lingüística de *Corpus* (BERBER SARDINHA, 2004), complementados por referenciais dos estudos lingüísticos de perspectiva enunciativa desenvolvidos por Émile Benveniste (1989), no recorte denominado Lingüística da Enunciação (FLORES, TEIXEIRA 2005). Partimos do pressuposto que o texto desse tipo de jornal integra um uso específico da

língua portuguesa, denominado aqui, provisoriamente, *Português Popular Escrito*. Por vocabulário, entenderemos o conjunto geral de palavras que conforma um dado uso da língua, sem distinção entre itens gramaticais e itens lexicais.

Naturalmente, é preciso considerar que o texto do jornal popular é construído por jornalistas, pessoas com nível cultural, econômico e social privilegiados em relação à grande massa da população brasileira. Esses jornalistas precisam, então, interagir com um tipo de leitor cujas condições são bem menos privilegiadas que suas, via de regra perpassadas por uma baixa escolaridade formal e baixo poder aquisitivo. Assim, em tese, no cenário do texto jornal popular, trata-se de um redator de formação universitária que normalmente maneja a escrita de um português culto, escrita cuja feição precisará ser adaptada para que haja a interação desejada entre o veículo, o jornalista e seu público-alvo.

Vale ressaltar, de antemão, que a grande maioria das pesquisas em *corpora* sobre vocabulário, sobre neologismos ou sobre outros elementos mórficos ou gramaticais da língua portuguesa, feitas no Brasil até hoje, não utilizam materiais desse gênero. Nossos pesquisadores e pesquisadores lusitanos, ao se ocuparem do português brasileiro escrito têm utilizado principalmente materiais oriundos do jornal *Folha de São Paulo* (KAUFMANN, 2008) e, em menor proporção, do jornal *O Estado de São Paulo* ou o jornal *O Globo* (MARONEZE, 2009). Em função dessa lacuna de dados lingüísticos sobre o vocabulário presente na escrita do jornalismo popular, em tese diferenciado do material jornalístico que usualmente se tem explorado, esta proposta de pesquisa também inclui a organização e oferecimento gratuito e *on-line* de um *corpus* específico para pesquisadores interessados.

Além desse *corpus*, outro objetivo da pesquisa é a obtenção de uma caracterização do vocabulário e da feição da linguagem tal como exibidos em textos que foram feitos, em tese, de um modo mais simplificado para atingir um público de baixo poder aquisitivo. Em uma primeira etapa, teremos como *corpus* de estudo uma amostra seriada de edições diárias dos 12 meses de 2008 do jornal popular *Diário Gaúcho* (doravante DG), esse material foi cedido a nós, com a devida autorização, emitida pelo editor-chefe do jornal, para armazenamento, compartilhamento e publicação *on-line* na versão somente texto.

A simplificação desse tipo de texto jornalístico, simplificação afirmada também em tese, dado que carece de confirmação empírica em um *corpus*, estaria a serviço de uma facilitação de compreensão de leitura para pessoas de um determinado grupo social e econômico, com uma bagagem cultural mais ou menos tipificada e graus de escolaridade relativamente baixos. A pesquisa visa, então, reconhecer empírica e estatisticamente quais são as suas características mais recorrentes no que refere às palavras e às associações de palavras mais empregadas.

Dada a carência de estudos lingüísticos sobre esse tipo de texto e a dimensão do *corpus*, o projeto que aqui se apresenta prevê a colaboração de outros pesquisadores de áreas conexas, tanto de colegas dos estudos da linguagem, que explorarão aspectos morfológicos, semânticos e enunciativos, além de aspectos relativos ao aproveitamento desse tipo de texto jornalístico para o ensino, quanto de colegas da área do Processamento da Linguagem Natural interessados em padrões de simplificação informação, representação de conteúdo e tratamento de expressões multipalavra em *corpora*. Cada um desses pesquisadores (vide a nominata em *Equipe de pesquisa*) deverá produzir estudos específicos sobre o corpus DG a partir dos seus diferentes pontos de interesse e perspectivas.

Os enfoques iniciais da pesquisa, na parte que cabe à coordenação, serão principalmente de cunho estatístico. Para esse tipo de abordagem do *corpus*, contaremos com o apoio de pesquisadores de Lingüística Computacional/Processamento da Linguagem Natural (PLN) do Instituto de Informática da UFRGS e da Faculdade de Informática (FACIN) da PUC-RS. Esses pesquisadores da área da Computação, Profa. Dra. Aline Villavicencio e Renata Vieira, desenvolvem estudos sobre sistemas de exploração automatizada de *corpora*,

sobre presença e configuração de expressões multipalavra e sobre sistemas automáticos para geração de ontologias a partir de *corpora*.

Prevemos também observações e contrastes com padrões de vocabulário do jornal *Zero Hora* (ZH), publicado pela mesma empresa do Diário Gaúcho, o qual é dirigido a públicos de maior poder aquisitivo. Esse contraste é realizado a partir de textos que são simultaneamente publicados no jornal ZH e no DG, sendo que a versão do texto do DG é, em geral, mais curta. Esse contraste DG-ZH será feito em parceria com investigadores de PLN do Núcleo Interinstitucional de Lingüística Computacional da USP (NILC-USP) que já desenvolvem pesquisa relacionada no projeto PorSimples[1]. Esse projeto, que tem apoio do CNPq, estuda a simplificação de textos com vista a atender portadores de dificuldades de leitura (MARGARIDO, PARDO, ALUISIO, 2009).

O *corpus* PorSimples inclui reportagens escritas do jornal ZH que possuem uma extensão denominada "Para seu Filho Ler". Essa extensão corresponde a uma versão simplificada do texto da reportagem dirigida para um público infantil. Nessa pesquisa, tem-se, assim, um corpus paralelo de "textos originais" e "textos adaptados para crianças de 8 a 11 anos", além de versões simplificadas dos originais destinadas a analfabetos funcionais e, potencialmente, a pessoas com outras deficiências cognitivas, como afasia e dislexia. Somaremos, a esse projeto, versões "populares" desses textos, estabelecendo, em tese, um segundo degrau de simplificação.

Também estão previstos outros contrastes com diferentes padrões de vocabulário: textos de revistas de divulgação de temas de ciências para leigos, redações de vestibulandos, textos científicos e textos literários. No segmento redações de evstibulandos, pretendemos aproveitar os resultados iniciais do trabalho de Finatto, Azeredo e Cremonese (2008) Para finalizar esta breve caracterização do nosso projeto, vale mencionar que imaginamos poder utilizar o *corpus* DG como uma referência para elaboração de um dicionário de português para estrangeiros, tendo em vista e feição, em tese, simplificada do texto e do seu vocabulário, descontados, naturalmente, aspectos regionais do vocabulário gaúcho possivelmente presentes no material.

Breve caracterização de referenciais teóricos

A pesquisa utiliza como principal referencial teórico a Lingüística de Corpus (doravante LC), tal como nos foi apresentada, no Brasil, por Berber Sardinha (2004). A LC será entendida neste trabalho como um tipo de abordagem teórica e metodológica dos estudos lingüísticos que privilegia o exame da linguagem em grandes conjuntos de textos autênticos, os *corpora*. Nela são investigadas as realizações lingüísticas possíveis e prováveis de serem produzidas por falantes reais e não por potenciais falantes idealizados.

Além disso, como a LC entende a língua como um sistema probabilístico de combinatórias, não se pode observar as palavras isoladas que conformam o vocabulário do texto do jornal popular. Isso não é possível porque, conforme Stubbs (2001, p. 3), o nosso conhecimento da linguagem e dos textos não se restringe a um conhecimento das palavras isoladas, mas é integrado fundamentalmente pelo conhecimento de combinatórias possíveis e pelo conhecimento cultural que essas combinatórias freqüentemente contêm. Cabe dizer ainda que a **LC** não deve ser definida como uma nova Lingüística, mas sim como uma nova via para a Lingüística (BERBER SARDINHA, 2004, p.35), visto que se ocupa, em meio a sua concepção peculiar de língua, da organização e da "mineração" de acervos textuais em formato digital.

[1] http://caravelas.icmc.usp.br/wiki/index.php/Principal

As bases teórico-metodológicas da Lingüística de Corpus devem-se aos trabalhos do britânico J.R. Firth (1980-1960) que, em um computador dos anos 50, já pesquisava em textos autênticos a distribuição de palavras sócio-culturalmente relevantes. Ele acreditava que o significado de uma palavra se configura no contexto de seu uso. Sua tão repetida citação "*You shall know a word by the company it keeps*" chama atenção para a imensa rede de relações sintagmáticas e paradigmáticas que envolve léxico e gramática, enfatizando o fenômeno que ele chama colocação. Observa, também que as palavras que o falante escolhe utilizar em meio a um todo de opções à sua disposição exibem um padrão de associação regular. Isto é, as palavras privilegiam um tipo de combinação ou, melhor dito, elas preferem determinadas associações e ainda rejeitam outras.

A **LC** vem dialogando mais intensamente com os estudos de Lexicologia e de Lexicografia, sobretudo fora do Brasil, desde os anos 80. E, essa aproximação, em termos do que vemos hoje no cenário brasileiro, deu-se, contudo, de um modo mais intenso apenas a partir dos anos 2000, em função do maior acesso da comunidade de pesquisa aos computadores e aos *softwares* para estatística lexical que contemplem o português brasileiro[2].

Outro motivo para a ampliação do reconhecimento da **LC**, no cenário brasileiro, foi a ênfase para a observação extensiva dos usos da língua em situações reais de comunicação, escritas ou orais. Essa ênfase pôde ser associada a uma das vocações da **LC**: o processamento extensivo de grandes *corpora* com o fim de identificar padrões de usos "reais" de língua. A língua, sob essa ótica, é entendida como um sistema probabilístico de combinatórias, de modo que uma dada palavra se define pela sua presença e pelos tipos de vinculação com as demais palavras dessa língua. Assim, para as palavras, também vale a máxima "me diga com quem andas e te direi quem és".

A evolução da Lingüística Textual e a afirmação da Lingüística de Corpus no campo dos estudos lingüísticos têm cada vez mais colocado o texto (ou conjuntos de textos, vistos agora como *corpus/corpora*) como um objeto central de estudo. Em decorrência disso, pesquisas voltadas para a identificação de características micro e macroestruturais nos e dos diferentes textos passaram as ser empreendidos para mostrar que, para além de aspectos formais mais pontuais, existem práticas discursivas, verdadeiros "modos de dizer" ou convencionalidades que são particulares de determinadas comunidades discursivas em determinados gêneros textuais. Esse assunto já foi extensivamente explorado por autores tais como Bakhtin (1988 e 1997), Swales (1990) e Marcuschi (2005 e 2006). Nessa medida, nosso trabalho também considerará o texto do jornal popular como um gênero ou macro-gênero textual, o qual deve exibir determinadas convenções de forma e de sentido.

Escolhemos o gênero jornal popular como foco de estudo por três motivos principais: **a)** são raras as pesquisa lingüísticas que os tomam como objeto de estudo ou descrição; **b)** sua configuração textual e lexical, em tese simplificadas, podem fornecer interessantes subsídios para estudos que se ocupem com os temas de sumarização ou de simplificação da linguagem e vocabulário; **c)** é grande a aceitação desse tipo de jornal por parte de seu público alvo em todo o Brasil, o que o torna um dos gêneros textuais de grande circulação e acesso da atualidade.

Sobre a noção de palavra e da sua observação no *corpus*

Palavras são unidades da língua e qualquer não-lingüista tem muita facilidade em identificar uma. Entre os lingüistas, entretanto, essa facilidade não se dá, pois há inúmeras concepções sobre o que seja uma palavra, e cada ponto de vista parece construir um objeto diferente. A terminologia da Lingüística justamente revela uma parte das diversidades envolvidas em sua conceituação: *palavra* também é denominada *item lexical*, *vocábulo formal ou mórfico*, *lexia*, *lema*, *forma lematizada*, *lexema***,** *item lexical***,** etc. E, como é fácil concluir,

[2] Até então, a maioria das ferramentas disponíveis estava adaptada apenas para reconhecer palavras em inglês.

cada nome corresponde a uma perspectiva de *palavra* que acompanha uma dada perspectiva do que seja língua e linguagem. Na prática lexicográfica, a palavra *manga* entendida como fruta tem um registro distinto – em verbete separado - da mesma *manga* de camisa... Itens homógrafos que passam a ser dicionarizados como duas palavras diferentes. Aqui o sentido define a diferença.

Conforme Biderman (1999, p.87) *vocábulo* e *palavra* são termos imprecisos. E, para minimizar as imprecisões, a autora propunha uma distinção entre *lexema* e *lema*. O primeiro é uma unidade virtual e abstrata que compõe o léxico[3]; o segundo é registro do lexema, de acordo com uma forma de referência, que é indicado em um dicionário.

Biderman (op.cit.) também nos apresenta o termo *lexia*, o qual entende como manifestação discursiva dos lexemas. Assim, as lexias são a face concreta dos lexemas, o seu uso da fala e escrita. As lexias, conforme a autora, podem ser simples, complexas ou compostas, dependendo do tipo de juntura ou não que apresentem. Um exemplo de lexia simples seria *pai*, para lexia complexa teríamoscomo exemplo *cesta básica*. À lexia composta correspondem as seqüências unidas por hífen como *pai-de-santo.* Observe-se aqui a ênfase para a escrita; há, naturalmente, outras condições para a fala.

Em que pesem tais distinções, na nossa pesquisa, o termo *palavra* corresponderá, muito singelamente, à concepção de *palavra gráfica*, entendida como um conjunto de caracteres entre dois espaços em branco. Assim, ***rsrsrsrs*** é tanto uma palavra tanto quanto ***cadeira*** e ***cadeiras*** ou ***de*** e ***dos***, não sendo, nesta primeira fase da nossa investigação, computadas as diferentes formas ou flexões de uma mesma base. Cada item gráfico será tomado como um item-palavra, independentemente de sua repetição ou variação. Isso é que se denomina, em LC, de *tokens* (itens) e que corresponde, grosso modo, ao número de palavras gráficas que há num texto. Essa perspectiva, bem sabemos, é bastante redutora, mas é uma opção metodológica necessária para o início de uma investigação.

Além disso, acreditamos que uma parte essencial e caracterizadora das palavras é a sua freqüência de uso. Consideraremos , assim, as freqüências das palavras no nosso *corpus* DG e das mesmas unidades em outros *corpora* do português brasileiro**.** Conforme depreendemos dos ensinamentos de Biderman (1998, p.162 ss.), que pioneiramente acompanhou referenciais da LC em seus trabalhos de lexicologia e lexicografia, a língua pode ser vista como um sistema probabilístico de combinatórias e de usos mais ou menos freqüentes, salientando a autora que a freqüência deve ser tomada como uma característica típica da palavra. Em uma perspectiva como essa, cada tipo de palavra registrada em um levantamento lexical terá padrões de freqüência que lhe são peculiares e isso deverá ser levado em conta na apresentação das informações sobre ela.

O léxico, o vocabulário, as palavras que, enfim, compõem uma língua, estão em toda parte e, ao mesmo tempo, associam-se a diferentes níveis da linguagem. Vemos as palavras sob diferentes perspectivas: pela ótica da fonologia, morfologia, sintaxe, semântica e também pela macroperspectiva do texto. Por isso, não seria demasiado supor que o léxico possa ser um grande intermediador entre os diferentes níveis de estruturação da língua.

A Lexicologia, procurando descrever e compreender os mistérios do vocabulário em meio às diversas interfaces da língua, no plano da gramática e/ou do discurso/texto, associa-se aos estudos de Lexicografia prática e Lexicografia teórica. Mais recentemente, no Brasil, vê-se uma aproximação entre Lexicologia e estudos baseados em *corpora*.
Como bem explica Marcuschi (2005), o léxico

> ao lado da sintaxe e da fonologia, (...) é o terceiro grande pilar da língua. Sem léxico não há língua. Mas o léxico é o nível da realização lingüística tido como o mais

[3] Não diferenciaremos aqui os termos *léxico* e *vocabulário*. Entretanto, essa distinção é útil em alguns momentos e no cenário de algumas oposições.

> instável, irregular e até certo ponto incontrolável. Podemos ver que tanto a sintaxe como a fonologia dispõem de um conjunto fechado de possibilidades básicas de realização numa língua, mas o léxico é aberto e todo dia presenciamos o surgimento de novos termos e o desaparecimento de antigos. Esta volatilidade não se dá por mero capricho dos falantes e das línguas e sim porque o léxico recebe usos sempre renovados (...) (MARCUSCHI, 2005, p. 6)

Ao ser dinâmico, instável e renovável, o léxico exigirá uma forma de tratamento igualmente dinâmica. O léxico é tão significativo e complexo que é capaz de identificar o falante, o gênero textual e a situação comunicativa. Por sua importância, natureza e magnitude, parece lógico que seu estudo seja feito em parcelas ou porções, de modo que, de várias frentes e ângulos, possamos colher dados para vislumbrar sua totalidade.

A concepção de palavra e de léxico que guiarão esta pesquisa procuram fazer uma integração entre uma visão estatística de **ocorrência**, associada à concepção de palavra gráfica, e uma visão enunciativa de **palavra**. A junção pode parecer estranha, mas, acreditamos, pode render bons resultados à medida que redimensiona a noção de uso. Afinal, cada ocorrência de uma dada palavra ao longo de um *corpus* corresponde a um uso. Se cada uso corresponde a um sentido, tem-se então uma infinidade de sentidos-usos. Entretanto, sabemos que a LC preconiza que se observem os diferentes usos para que então se possa descrever seus padrões e combinatórias, homogeneidades e heterogeneidades, o que nos dará uma idéia de uma *prosódia semântica* de uma dada palavra ou construção em meio às suas diferentes apresentações de uso.

Segundo Benveniste *(*1989), há dois sistemas na linguagem: o semiótico, das formas, cuja unidade é o **signo**, e o semântico, do uso, do discurso, cuja unidade é a **palavra**[4]. O sujeito toma os signos, do sistema semiótico, e os significa, dando-lhes um sentido, tornando-os palavras da frase (estando, portanto, no nível semântico), de acordo com a instância desse discurso. Assim, o **signo** torna-se **palavra** pela enunciação.

Ao mesmo tempo em que constitui a própria instância do discurso, o sujeito constitui a si mesmo como "eu" – "o ato individual de apropriação da língua introduz aquele que fala em sua fala" (1989, p.84) –, constitui também um "tu" – elemento sem o qual não há linguagem. Afinal, a subjetividade (a noção mesma de sujeito) apenas pode ser alcançada por contraste, pela intersubjetividade (é apenas em relação a um outro que o homem consegue se constituir como sujeito).

É aqui, portanto, que se coloca a questão lexical na teoria benvenistiana, a partir da noção de **palavra**. Uma *palavra* para Benveniste não é, portanto, apenas um elemento físico de uma frase, mas uma *ocorrência* cuja referenciação – interna à própria enunciação e sem relação com o mundo físico – determina seu sentido, dependendo do *eu-tu-ele-aqui-agora* que decorre da apropriação da língua pelo sujeito.

As **palavras** estão em toda parte e permeiam, peculiarmente, as diferentes dimensões da linguagem. Respeitando o acima citado sobre a distinção entre o que tange à semiótica e o que tange ao discurso-uso, teríamos uma perspectiva de vocabulário que é conformada pelo plano da "palavra-signo" e pelo plano da "palavra-em-uso". Apenas desse viés enunciativo e de tantos outros que se quisesse evocar, já vemos que é possível perceber e tratar as palavras que perfazem uma língua sob diferentes ângulos: fonologia, morfologia, sintaxe, semântica, texto, discurso e enunciação. O léxico, elemento que será aqui destacado no cenário específico do jornal popular, atua como um grande intermediador entre os diferentes planos da língua,

[4] Vale destacar que a separação dos níveis semiótico e semântico somente pode ser feita para fins didáticos. Além disso, tais níveis não podem ser identificados com as noções de língua e fala de Saussure. Embora a teoria benvenistiana seja intrinsecamente relacionada à teoria saussuriana, Benveniste vai além do que propunha Saussure, uma vez que a noção de nível semântico engloba o uso da língua pelo sujeito.

planos que os diferentes pontos de vista de descrição e de investigação lingüística podem instaurar.

Conforme já mencionado, um dos métodos de observação que adotamos para o estudo do nosso *corpus* de jornais populares é oriundo da **LC**. A **LC** é uma perspectiva diferenciada dos estudos da linguagem, bastante marcada pela observação, o mais extensiva possível, dos usos da língua e pelo apoio de recursos informatizados aplicados a acervos textuais em formato digital. Esses acervos, os *corpora*, são especialmente organizados para a pesquisa lingüística e devem permitir demonstrar padrões e especificidades dos usos da língua em diferentes situações. A partir disso, nossa intenção é tentar descobrir que padrões de uso de palavras estariam postos no material oriundo do DG.

Neste trabalho, conforme a **LC**, entenderemos a noção de **riqueza lexical** como uma medida estatística, como uma relação que se estabelece entre o número de palavras que perfaz um texto e o número de palavras repetidas e diferentes nele encontradas. Trata-se, assim, apenas de uma relação entre o número total de ocorrências (*tokens*) e o número de formas diferentes (*types*). Quanto maior for o número de *types*, maior será a riqueza e a variedade do vocabulário. Quanto menor o percentual[5], mais repetitivo é o universo de vocabulário. Outro elemento importante a considerar no universo da riqueza lexical, assim entendida, ainda que restrito ao número de palavras e à sua variedade, é o número de palavras cuja ocorrência é única em um corpus. Essas palavras são denominadas *hapax legomena* ou hápax. Interessa, enfim, também observar que palavras ocorrem apenas uma vez ao longo de um dado texto para que se componha um retrato mais abrangente de seu universo vocabular.

Bastante longe de considerar a língua sob um ponto de vista estatístico, uma visão enunciativa – à medida que considera que cada evento enunciativo é único, irrepetível e auto-referenciado lingüisticamente – vai analisar caso a caso, ocorrência a ocorrência, sem perder de vista que qualquer nível da língua – como o fonológico, o morfológico e o sintático, por exemplo –, bem como quaisquer de seus aspectos, estão sempre subordinados ao sentido.

Dessa maneira, uma abordagem apenas quantitativa não poderia ser produzida por um estudo enunciativo, posto que o importante nessa perspectiva é a singularidade do processo de enunciação. A visão enunciativa complementará, portanto a observação estatística da LC à medida que nos permitirá ponderar, em diferentes condições, sobre as ocorrências, sejam elas *tokens* ou *types*, e oferecerá um leitura peculiar para o que observarmos na *linhas de concordância* que espelham os contextos de uso de uma dada palavra. Há, entretanto, que se considerar que os pontos singulares inegavelmente desvelados pelas abordagens estatísticas, ao mostrarem singularidades em meio a padrões, podem sinalizar ricos filões de exploração para a análise em moldes benvenistianos.

Não é demais supor que um enfoque começa onde o outro oferece, em tese, um ponto-cego. Para a **LC**, estão as palavras e as suas companhias, as suas reiterações, as suas diversidades, enquanto que, para a Enunciação, estão as pessoas e suas escolhas de significação, os efeitos de sentido e os seus modos de dizer em meio às diferentes possibilidades que a escrita do jornalismo popular oferece.

Materiais e métodos

Nesta primeira etapa da pesquisa (fase1- em 24 meses), será estudado apenas um jornal diário que corresponde ao tipo *jornal popular*. Entre várias publicações do gênero disponíveis no Brasil, optamos pelo jornal **Diário Gaúcho** (doravante **DG**), publicado em Porto Alegre-RS, produzido pelo grupo RBS. Esse jornal foi selecionado entre outros similares disponíveis no Brasil em função dos seguintes fatores: **a)** estar em circulação já há 10 anos; **b)** ter alta tiragem/dia – sua tiragem média atual é de 145 mil exemplares/dia; **c)** já

[5] Esse percentual é a razão *type/token* (forma/ocorrência).

ter sido estudado em trabalhos e pesquisas da área do Jornalismo (AMARAL, 2004 e 2006; BERNADES, 2004); **d)** contarmos com a autorização para armazenagem e compartilhamento *on-line* do *corpus* por parte do próprio jornal.

O jornal **DG,** como já mencionado, tem grande tiragem, salientando-se que dados do IVC (Instituto Verificador de Circulação) indicam que cada 01 exemplar tende a ser lido por 05 pessoas em média, fato que redimensiona a relação entre tiragem/venda/número de leitores. É o único do gênero publicado na cidade de Porto Alegre e sua tiragem atesta a grande aceitação por parte de seu público-alvo na cidade e em todo um entorno de municípios vizinhos da região metropolitana. O seu número de leitores supera, de longe, o de jornais da cidade de Porto Alegre dirigidos a públicos mais tradicionais ou de maior poder aquisitivo e que são distribuídos em todo o Estado do Rio Grande do Sul. Isso aplica, por exemplo, ao tradicional jornal *Zero Hora*, produzido há 35 anos pelo mesmo grupo empresarial do **DG,** cujo público leitor em todo o estado do RS parece ter se tornado pequeno frente ao número de leitores do DG apenas na grande Porto Alegre.

Outro fator digno de nota é a grande adesão de seu público a quaisquer eventos promovidos pelo jornal. Isso demonstra, em tese, uma eficiente relação mercadológica, provavelmente acompanhada de uma metodologia bem-sucedida de elaboração de texto dirigido para um tipo de leitor determinado. Além disso, conforme Amaral (AMARAL, 2006, p.80), "parte dos consumidores do DG não eram leitores de jornal, e após seu lançamento, em 2000, a região metropolitana de Porto Alegre passou a ser a primeira em índice de leitura de jornais no Brasil."

No que se refere à coleta e seleção de amostra das edições diárias do DG para a organização do *corpus* estudo, foi utilizada a mesma sistemática adotada por Ieda Maria Alves no Projeto TermNEO, acessível em http://www.fflch.usp.br/dlcv/neo que corresponde à iniciativa Projeto "Observatório de Neologismos Científicos e Técnicos do Português Contemporâneo" e no seu Projeto "**Base de Neologismos do Português Brasileiro Contemporâneo"**, que utilizam, desde 1993, textos de jornais diários e de revistas semanais como *corpus* (MARONEZE, 2009, p.531)

Tendo sido tomados todos os jornais do ano de 2008, foram a cada mês, selecionados de 10 a 12 dias em cada mês distribuídos nas 4 ou 5 semanas de cada mês. A intenção foi obter-se uma amostra composta pelo todo do jornal de dada dia (excluídos apenas informes publicitários, classificados, indicações de expediente e datação) por diferentes dias não consecutivos de cada semana. Como o DG circula de segunda a sábado, assim, em janeiro de 2008, na primeira semana, foram selecionados as edições de segunda, quarta e sexta-feira. Na segunda semana de janeiro, as edições de terça, quinta e sábado e assim sucessivamente. Esse procedimento foi aplicado aos 12 meses de 2008.

Além desse material, compusemos também uma pequena amostra com todos os dias da primeira semana de janeiro de 2009, uma amostra-controle das edições de segunda a sábado. Todos os arquivos cedidos pelo jornal DG estão em fomato PDF, sendo que cada arquivo corresponde a cada umas das páginas do jornal.

Cada edição diária do jornal tem em média 32 páginas, as quais exibem abundância de fotos e ilustrações acompanhados por textos relativamente curtos. É importante salientar também que o DG publica algumas reportagens produzidas pelo jornal Zero Hora, que pertence à mesma empresa. A feição dos textos é, via de regra, reduzida ou simplificada havendo um maior enriquecimento de imagens.

Em função dessa transposição de textos entre veículos diferentes de uma mesma empresa, pareceu-nos importante conjugar nosso estudo do DG aos estudos do jornal ZH em andamento pela Profa. Dra. Sandra Aluísio do NILC-SP. Essa conjugação visa tentar verificar como teriam sido operadas essas simplificações, as quais adaptam o texto do veículo dirigido a um público de maior poder aquisitivo para consumo do leitor de menor poder aquisitivo.

Quanto à metodologia de trabalho com a pesquisa como um todo e com esse *corpus* do *DG*, em função do acima exposto, prevemos diferentes etapas. Entretanto, a primeira e mais imediata tarefa de pesquisa consiste na produção e organização do *corpus* **DG** em formato somente texto, o que é feito a partir dos arquivos originais a nós cedidos em formato PDF. Isso é necessário porque apenas nesse formato de arquivo é possível utilizar sistemas computacionais especialmente desenvolvidos para a realização de diversas estatísticas lexicais, para a produção de representações de conteúdos textuais e para a obtenção de listagens de palavras de diferentes perfis. Concomitantemente a essa tarefa, temos as seguintes etapas gerais e específicas.

Etapas gerais:

1) revisão da literatura sobre jornalismo popular e tratamento estatístico do léxico em *corpora* oriundos de jornais diários;
2) registro e sistematização das características textuais de textos de jornais populares reconhecidas na bibliografia de Jornalismo;
3) registro e sistematização das características textuais de textos de jornais convencionais reconhecidas na bibliografia de PLN e de Lingüística de Corpus;
4) revisão da literatura sobre Lexicologia, incluindo pesquisas sobre neologismos, amparada em observações em *corpora* de jornais produzidos no Brasil*;*
5) composição de pequena amostra com uma edição de cada dia de diferentes jornais populares publicados em diferentes estados do Brasil;
6) sistematização das características lexicais e textuais da amostra cima citada.

Etapas específicas com o *corpus* DG (edições de 12 meses de 2008):

1) conversão de cada página PDF do DG da edição de determinado dia para o formato TXT (somente texto);
2) exclusão de texto publicitários, datações e de textos de expediente – cuja presença seja fixa em todos os dias;
3) junção em arquivo único dos arquivos de cada dia do jornal;
4) junção em arquivo único dos arquivos de cada mês do jornal;
5) conferência com apoio informatizado e conferência manual de arquivos de cada mês para eliminação de caracteres indesejáveis, gerados pela conversão;
6) geração de listas de freqüência de palavras por cada mês com ferramentas disponíveis gratuitamente *on-line* ;
7) geração de listas de freqüências de construções recorrentes (*clusters*) –idem acima para ferramentas gratuitas *on-line*;
8) geração de dados estatísticos gerais sobre as palavras utilizadas por de cada mês do DG, incluindo a observação de número de palavras de número de formas diferentes de palavras (*tokens* e *types)*, observação de tamanho de palavras por caracteres e de tamanho de sentenças por número de palavras (estas últimas feitas com apoio do *software* Wordsmith tools v.3.0).
9) comparação de dados mensais do DG em relação a dados do jornal Zero Hora nas edições do mesmos dias gerados pelo Projeto Porsimples da Profa. Sandra Aluísio do NILC/SP;
10) observação contrastiva pontual de textos transpostos do jornal Zero Hora para o DG;
11) comparação dos dados sobre vocabulário mais recorrente do DG (incluindo contraste de itens de freqüência 1) com dados de *corpora* integrados pelos jornal Folha de São Paulo oferecidos pelo NILC/Linguateca e Banco de Português;
12) geração de caracterizações sobre o vocabulário mais recorrente e vocabulário peculiar do DG;

13) produção de arquivos para publicação *on-line* do corpus;
14) compartilhamento de informações e do corpus reunido com os pesquisadores-colaboradores e seus orientandos;
15) publicação *on-line* de dados da pesquisa;
16) apresentação de trabalhos em eventos e publicações de artigos.

Produtos da pesquisa

Os produtos mais imediatos da pesquisa, já ao final do seu primeiro ano, serão:

a) 50% do *corpus* reunido e devidamente disponível *on-line*, acompanhado de suas informações estatísticas básicas (listas de freqüência de palavras e de construções recorrentes, relação entre número de formas diferentes e número de palavras em cada mês do jornal);
b) descrições da configuração do vocabulário manifestado em segmentos mensais do *corpus* reunido.
c) listagens contextualizadas de itens e de expressões mais recorrentes do vocabulário do *corpus* DG por segmentos mensais;
d) estudos contrastivos com vocabulários de outros *corpora.*
e) estudos-piloto sobre vocabulário do português popular escrito do Brasil com vistas à produção de dicionários básicos de português de acordo de freqüências de usos;
f) trabalhos para apresentação em eventos científicos das áreas de Lingüística/Letras, de Lingüística de Corpus e de Lingüística Computacional;
g) um catálogo inicial *on-line* com as palavras e as construções mais utilizadas no DG, acompanhadas de seus contextos e de comentários, como também um levantamento sobre padrões de frase mais recorrentes. Cada registro desse catálogo será acompanhado de suas freqüências relativas no *corpus* sob exame e de algumas informações contextuais.

Ao final de 24 meses prevemos:

a) oferecimento de 100% do corpus DG devidamente etiquetado e identificado, para ser utilizado por pesquisadores universitários interessados no tema;
b) um catálogo completo *on-line* com as palavras e as construções mais utilizadas no *corpus* DG, acompanhadas de seus contextos e de comentários, como também um levantamento sobre padrões de frase mais recorrentes

Estão previstos, assim, como produtos continuados da pesquisa ao longo de seus 24 meses, estudos-piloto sobre determinados tipos de palavras ou construções, publicação de artigos e apresentação de trabalhos em eventos pelo diferentes pesquisadores-colaboradores e pela pesquisadora responsável. Nesses eventos, serão divulgados dados da exploração em suas diferentes etapas. A idéia da produção do conhecimento sobre o *corpus* do DG segue a lógica da produção de dados partindo da ótica da LC, os quais são oferecidos para as diferentes explorações dos pesquisadores-colaboradores.

Como produtos continuados desta investigação, prevemos também a formação em iniciação científica de estudantes de graduação e o aproveitamento do *corpus* reunido e de suas descrições por parte de orientandos de mestrado e de doutorado dos pesquisadores-colaboradores envolvidos.

Equipe de trabalho inicial prevista: 04 estudantes de graduação do curso de Letras da UFRGS, 01 estudante de Ciência da Computação, 01 mestrando e 01 doutorando do Programa de Pó-Graduação em Letras da UFRGS.

Prazo de execução da pesquisa:
fase 1 – organização do *corpus* e estudos iniciais, oferecimento parcial on-line do corpus - 12 meses- primeiro ano
fase 2: oferecimento *on-line* do total do corpus e de levantamentos de padrões do vocabulário por tipos de itens e por tipos de freqüências - 12 meses –segundo ano.

Local da realização da pesquisa: UFRGS, Instituto de Letras, Porto Alegre - RS. Estudos a partir do corpus DG também serão realizados na PUC-RS, UNISINOS, UERGS e NILC-USP.

Cronograma
Tarefas
Etapas gerais:

1) revisão da literatura sobre jornalismo popular e tratamento estatístico do léxico em *corpora* oriundos de jornais diários;
2) registro e sistematização das características textuais de textos de jornais populares reconhecidas na bibliografia de Jornalismo;
3) registro e sistematização das características textuais de textos de jornais convencionais reconhecidas na bibliografia de PLN e de Lingüística de Corpus;
4) revisão da literatura sobre Lexicologia, incluindo pesquisas sobre neologismos, amparada em observações em *corpora* de jornais produzidos no Brasil*;*
5) composição de pequena amostra com uma edição de cada dia de diferentes jornais populares publicados em diferentes estados do Brasil;
6) sistematização das características lexicais e textuais da amostra cima citada.

Etapas específicas com o *corpus* DG

7) conversão de cada uma das páginas PDF do DG da edição de determinado dia do mês para o formato TXT (somente texto);
8) exclusão de texto publicitários, datações e de textos de expediente – cuja presença seja fixa em todos os dias;
9) junção em arquivo único dos arquivos de cada dia do jornal;
10) junção em arquivo único dos arquivos de cada mês do jornal;
11) conferência com apoio informatizado e conferência manual de arquivos de cada mês para eliminação de caracteres indesejáveis, gerados pela conversão;
12) geração de listas de freqüência de palavras por cada mês com ferramentas disponíveis gratuitamente *on-line* ;
13) geração de listas de freqüências de construções recorrentes (*clusters*) –idem acima para ferramentas gratuitas *on-line*;
14) geração de dados estatísticos gerais sobre as palavras utilizadas por de cada mês do DG, incluindo a observação de número de palavras de número de formas diferentes de palavras (*tokens* e *types)*, observação de tamanho de palavras por caracteres e de tamanho de sentenças por número de palavras (estas últimas feitas com apoio do *software* Wordsmith tools v.3.0);
15) comparação de dados mensais do DG em relação a dados do jornal Zero Hora nas edições do mesmos dias gerados pelo Projeto PorSimples da Profa. Sandra Aluísio do NILC/SP;

16) observação contrastiva pontual de textos transpostos do jornal Zero Hora para o DG;
17) comparação dos dados sobre vocabulário mais recorrente do DG (incluindo contraste de itens de freqüência 1) com dados de *corpora* integrados pelos jornal Folha de São Paulo oferecidos pelo NILC/Linguateca e Banco de Português;
18) geração de caracterizações sobre o vocabulário mais recorrente e vocabulário peculiar do DG;
19) produção de arquivos para publicação *on-line* do corpus;
20) compartilhamento de informações e do corpus reunido com os pesquisadores-colaboradores e seus orientandos;
21) publicação *on-line* de dados da pesquisa;
22) apresentação de trabalhos em eventos e publicações de artigos.

Quadro de meses e de tarefas

3.2 CRONOGRAMA PREVISTO: 2006 A 2008

TAREFAS	**12 Meses – primeiro ano**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	X	X	X	X	X	X						
2	X	X	X	X	X	X						
3			X	X	X	X	X	X				
4	X	X	X	X	X	X						
5						X	X	X	X	X		
6						X	X	X	X	X	X	
7		X	X	X	X	X	X	X	X	X	X	X
8		X	X	X	X	X	X	X	X	X		
9		X	X	X	X	X	X	X	X	X		
10					X	X	X	X	X			
11					X	X	X	X	X	X		
12						X	X	X	X	X	X	X

TAREFAS	**12 Meses – segundo ano**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
13	X	X	X	X	X							
14					X	X	X	X	X	X	X	
15					X	X	X	X	X	X		
16						X	X	X	X	X		
17						X	X	X	X	X		
18						X	X	X	X	X	X	X
19					X	X	X	X	X	X	X	X
20-21-22	X	X	X	X	X	X	X	X	X	X	X	X

Disponibilidade de infra-estrutura e de apoio técnico para o desenvolvimento do projeto

No gabinete de pesquisa da proponente, embora com espaço físico acanhado, há computadores disponíveis para o trabalho com o *corpus*. Esses computadores foram adquiridos graças à participação da pesquisadora proponente em Editais do CNPq, Edital Universal 2006 e Edital de Ciências Humanas e Sociais 2005.

Recursos financeiros de outras fontes

A pesquisa a ser empreendida já conta com apoio da SEAD, Secretaria de Educação a Distância da UFRGS, que nos concedeu uma bolsista de graduação da área de Informática para nos auxiliar no trabalho inicial de organização do *corpus* DG e do seu oferecimento inicial *on-line*. Este trabalho já está sendo feito atualmente e cobre o primeiro trimestre do *corpus*. Trata-se de uma iniciativa que visa integrar a pesquisa sobre vocabulário às atividades da disciplina Léxico e Dicionários, do curso de graduação em Letras da UFRGS. Essa disciplina tem 20% das suas atividades a distância.

Parcerias de estudo e pesquisa:

Os pesquisadores-colaboradores da equipe desta proposta de investigação já são parceiros de publicação e de realização de trabalhos de pesquisa da proponente, conforme pode ser verificado no seu CV Lattes. A profa. Dra. Sandra Aluísio do NILC/USP já colabora com a pesquisa com o *corpus* do DG para fins de Educação a Distância com apoio da SEAD-UFRGS. Ver em http://www.ufrgs.br/sead, Editais e Bolsas, Edital 11, Projeto Contemplados, o Projeto *PERFIS DO PORTUGUÊS POPULAR ESCRITO PARA UM AMBIENTE DE EAD: PADRÕES DO VOCABULÁRIO DE JORNAIS POPULARES BRASILEIROS PARA O ENSINO DE LÍNGUA PORTUGUESA*. http://www6.ufrgs.br/sead1/edital11/adm/exibir.php?id=127

O mesmo se aplica às Profas. Dras. Aline Villavicencio e Renata Vieira, Carmem Luci da Costa Silva e ao Prof. Dr. Valdir do Nascimento Flores.

Recursos solicitados:

R$ 33.000,00 (trinta e um mil reais) para 24 meses, recurso que será divididos em:
Serviços de terceiros e material de consumo: R$ 14.000,00; Equipamentos: R$ 8.000,00; Passagens para participação em eventos no país: R$ 7.000,00; diárias para participação em eventos R$ 3.000,00; Material bibliográfico R$ 1.000,00.

Bibliografia citada

AMARAL, Márcia Franz. Lugares de fala do leitor no Diário Gaúcho Universidade Federal do Rio Grande do Sul. Faculdade de Biblioteconomia e Comunicação. Programa de Pós-Graduação em Comunicação e Informação, Tese de Doutorado, 2004.

AMARAL, Márcia Franz. *Jornalismo Popular*. São Paulo: Contexto, 2006.

BERNARDES, Cristiane Brum. As Condições de produção do jornalismo popular massivo: o caso do Diário Gaúcho. Universidade Federal do Rio Grande do Sul. Faculdade de Biblioteconomia e Comunicação. Programa de Pós-Graduação em Comunicação e Informação. Diss. Mestrado, 2004.

BAKHTIN, M. (1988) *Marxismo e filosofia da linguagem*. São Paulo: Hucitec, 4ª Ed. Traduzido por Michel Lahud e Yara Frateschi Vieira a partir da edição francesa.

BAKHTIN, M. (1997) *Estética da criação verbal*. São Paulo: Martins Fontes, 2ª ed. 1ª ed. Em russo: 1979. Traduzido do francês por Maria Ermantina Galvão G. Pereira.

BERBER SARDINHA, T. *Lingüística de Corpus*. São Paulo: Manole, 2004.

BENVENISTE, Émile. *Problemas de lingüística geral II*. Campinas: Pontes, Editora da UNICAMP, 1989.

BIDERMAN, M. T. Conceito lingüístico de palavra. In: BASILIO, M. (org) *Palavra*. Rio de Janeiro: Grypho, 1999. vol.1, p.81-97.

BIDERMAN, M. T. A face quantitativa da linguagem: um dicionário de freqüências do português. *Alfa*, São Paulo, v.42 (esp.), p.161-181, 1998.

FINATTO, M.B.F. ;CREMONESE, L.E.; AZEREDO, S. O vocabulário na redação de vestibular: do enfoque estatístico às especificidades da enunciação. In: ABREU, S. (org*.) A redação no vestibular: do leitor ao produtor de texto*. COPERSE/UFRGS. Porto Alegre: Editora da UFRGS, 2008, p.95-108.

KAUFFMANN, Carlos H. . Elementos para uma análise quantitativa da linguagem do jornal. In: Stella Esther Ortweiler Tagnin; Oto Araújo Vale. (Org.). *Avanços da Lingüística de Corpus no Brasil*. 1 ed. São Paulo: Humanitas, 2008, v. 1, p. 407-418.

MARCUSCHI, Luiz Antônio. (2005a) "Gêneros textuais: definição e funcionalidade". In: *Gêneros textuais & ensino*. DIONISIO, Angela Paiva; et al. (Orgs.). 4ª ed. rev. e ampl. Rio de Janeiro: Lucerna, 2005. p. 19-36.

MARCUSCHI, Luiz Antônio. (2005b) *O Léxico: Lista, Rede ou Cognição Social?* (2005). Texto inédito, reformulado a partir da versão apresentada no V CICLO DE SEMINÁRIOS EM PSICOLOGIA COGNITIVA *COGNIÇÃO E LINGUAGEM,* da Universidade Federal de Pernambuco, Pós-Graduação em Psicologia Cognitiva, Recife, de 2 a 4 de dezembro de 2003.

MARCUSCHI, Luiz Antônio. "Gêneros textuais: configuração, dinamicidade e circulação". In: *Gêneros textuais – reflexões e ensino*. KARWOSKI, Acir Mário; et al. (Orgs.). 2ª ed. rev. e ampl. Rio de Janeiro: Lucerna, 2006. p. 23-36.

MARGARIDO, Paulo R. A., PARDO, Thiago A. S. e ALUÍSIO, Sandra M. Sumarização Automática para Simplificação de Textos: Experimentos e Lições Aprendidas " In: MELO, A. M., PICCOLO, L. S. G., ÁVILA, I. M. A, TAMBASCIA, C. A. (Org.). Usabilidade, Acessibilidade e Inteligibilidade Aplicadas em Interfaces para Analfabetos, Idosos e Pessoas com Deficiência: Resultados do Workshop. Campinas: CPqD, 2009. 73p. Disponível em: <http://www.cpqd.com.br/file.upload/1749021822/resultados_workshop_uai.pdf>, pp. 63-71.

MARONEZE, B. O. Adjetivos neológicos em um corpus da imprensa brasileira contemporânea. In: VI Congresso Internacional da Abralin, 2009, João Pessoa. Anais do VI Congresso Internacional da Abralin. João Pessoa : Ideia, 2009. p. 530-538.

STUBBS, M. (2001). Words in use: introductory examples. In: *Words and phrases. Corpus studies of lexical semantics*. Oxford: Blackwell, 2001. p. 3-23.

SWALES, J.M. *Genre analysis*: English in academic and research settings. Cambridge: Cambridge University Press, 1990.

FLORES, Valdir do Nascimento; TEIXEIRA, Marlene. *Introdução à Lingüística da Enunciação*. São Paulo: Contexto, 2005.