http://dx.doi.org/10.5007/1984-8420.2013v14n2p156

BANCO DE DADOS *FALARES SERGIPANOS*

FALARES SERGIPANOS DATABASE

Raquel Meister Ko. Freitag
Professora do Departamento de Letras Vernáculas
Programa de Pós-Graduação em Letras da Universidade Federal de Sergipe
rkofreitag@uol.com.br

RESUMO: Apresentamos a metodologia e os pressupostos norteadores da constituição da amostra sincrônica do banco de dados *Falares Sergipanos*, base de dados linguísticos ampla da variedade de português do Estado de Sergipe, seguindo duas linhas de coleta – a de comunidades de fala (estratificação homogeneizada) e a de comunidades de práticas (relações sociopessoais). Seu propósito é dar subsídios à investigação de variedades linguísticas do português, em seus diferentes níveis (do morfofonológico ao discursivo) e com diferentes propósitos (dos descritivos aos moldes sociolinguísticos às aplicações educacionais, subsidiando programas de ensino de língua materna). Partindo de estudos precedentes de mapeamento linguístico em Sergipe, são apontados os encaminhamentos para a estratificação da amostra e explicitados os procedimentos de coleta, armazenamento e disponibilização das amostras que compõem o banco de dados *Falares Sergipanos*.
PALAVRAS-CHAVE: Sociolinguística; Comunidade de fala; Comunidade de práticas.

ABSTRACT: In this paper we present the methodology and the theoretical guidelines to the development of Falares Sergipanos, *a synchronic linguistic database of Portuguese from the state of Sergipe, following two collecting approaches – speech communities (homogeneous stratification) and practice communities (socio-personal relationships). Its purpose is to subside the investigation of different linguistic varieties from Brazilian Portuguese, in different levels (from phonology to discourse) and with different applications (from sociolinguistic to education studies, subsiding language teaching). Based on previous studies of Sergipe's linguistic mapping, the text indicates aspects concerning the stratification, collecting procedures, storage and sharing of the samples that form the database.*
KEYWORDS: *Sociolinguistics; Speech community; Practice community.*

INTRODUÇÃO

A constituição e/ou ampliação de bancos de dados sociolinguísticos contemplando uma variedade do português brasileiro ainda não mapeada (ou pouco mapeada), como é o caso de Sergipe, é altamente desejável, motivo que levou à constituição de um banco de dados sociolinguísticos e diacrônicos do falar sergipano. Com o objetivo de subsidiar a investigação de variedades linguísticas do português, em seus diferentes níveis (do morfofonológico ao discursivo) e com diferentes propósitos (dos descritivos aos moldes sociolinguísticos às aplicações educacionais, subsidiando programas de ensino de língua

materna), o projeto *Falares Sergipanos* visa a constituir um banco de dados linguísticos amplo, abarcando as perspectivas sociolinguística (dados sincrônicos) e histórica (dados diacrônicos), da variedade de português do Estado de Sergipe. Nesta primeira etapa de desenvolvimento do projeto, a ênfase é dada à dimensão sincrônica do banco de dados.

A estrutura deste texto é a seguinte: inicialmente, são apresentadas possibilidades de uso de bancos de dados linguísticos, de maneira geral; no segundo momento, apresentam-se projetos precedentes de mapeamento linguístico em Sergipe, apontando os encaminhamentos para a estratificação da amostra; e, por fim, são explicitados os procedimentos de coleta, armazenamento e disponibilização dos dados linguísticos que compõem o banco de dados *Falares Sergipanos*.

2 BANCOS DE DADOS LINGUÍSTICOS: USOS

A pesquisa linguística baseada em usos requer a constituição de *corpus*, tarefa que é dispendiosa não só quanto a recursos financeiros, mas também quanto ao tempo. Iniciativas para viabilizar estudos linguísticos baseados em usos otimizando os recursos (muito escassos) – ao invés de cada pesquisador realizar a sua própria coleta de dados para a sua investigação – têm se tornado prática no cenário nacional, com a constituição de bancos de dados de acordo com as perspectivas sociolinguística e geolinguística, nas dimensões sincrônicas e diacrônicas, e também da linguística de *corpus*. Com cada projeto constituindo seu banco de dados em uma dada comunidade de fala, o mapeamento das variedades do português no Brasil vai tendo condições de ser efetivado: Salomão (2011) apresenta um panorama da Sociolinguística no Brasil, levantando 48 grupos de pesquisa que estão atuando nesta área do campo de estudos da linguagem.

O cenário dos estudos linguísticos contemporâneos aponta que hoje não podemos falar de A língua portuguesa, mas de AS línguas portuguesas. Para contemplar a pluralidade de comportamentos linguísticos, faz-se necessário investir na descrição de variedades. No entanto, diferentemente do ALiB – Atlas Linguístico do Brasil – e do PHPB – Projeto para a História do Português Brasileiro –, que são projetos nacionais da área de Dialetologia e Sociolinguística (o primeiro com alinhamento à Geolinguística e o segundo à Linguística Histórica), Gonçalves (2012) destaca que a Sociolinguística Variacionista não tem um grande projeto nacional unificado. Para garantir uma unidade à descrição provida pelo aparato da Sociolinguística Variacionista no Brasil, a padronização dos procedimentos metodológicos permite, posteriormente, a realização de estudos contrastivos entre as variedades, o que viabiliza uma descrição ampla do português brasileiro e o estabelecimento de suas normas. Trabalhos nesta direção têm sido aventados nos últimos anos no cenário nacional (cf. FREITAG; MARTINS, TAVARES, 2012).

A coleta de dados reais não é tarefa fácil, nem rápida. Por isso, bancos de dados linguísticos, sejam os constituídos nos moldes variacionistas, geolinguísticos ou diacrônicos, costumam ser utilizados para a pesquisa de diversos fenômenos de variação e

http://dx.doi.org/10.5007/1984-8420.2013v14n2p156

de mudança linguística, e alguns já são disponibilizados na internet, com amostras de áudio e de transcrição de dados.[1]

Há que se ressaltar que a sociolinguística variacionista no cenário nacional é uma das áreas que mais tem se desenvolvido. Outros projetos de descrição do português têm sido implementados e, consequentemente, outros bancos de dados vêm sendo constituídos, o que significa mais oportunidades de experiências com a heterogeneidade linguística do Brasil à disposição do professor de Língua Portuguesa.

A utilidade dos bancos de dados, entretanto, vai além dos estudos descritivos da língua. Trata-se de uma fonte rica a ser explorada nas aulas de Língua Portuguesa, permitindo que o professor realize atividades que propiciem ao aluno a experiência com a heterogeneidade linguística da comunidade de fala brasileira, no tocante à variação diatópica: será que um sergipano fala como um catarinense? E será que um sergipano da capital fala como um sergipano do interior? Os Parâmetros Curriculares Nacionais de Língua Portuguesa (BRASIL, 1998) preconizam este enfoque para o ensino de língua materna voltado para a diversidade e a variedade. Assim, bancos de dados linguísticos transformam-se em um recurso didático (embora sua finalidade primeira não seja esta) disponível gratuitamente e de fácil uso por parte do professor de Língua Portuguesa, além de possibilitar a incorporação de novas tecnologias ao ensino.

3 PRESSUPOSTOS NORTEADORES DA COLETA

Assumindo a perspectiva de que a Sociolinguística é o campo dos estudos linguísticos que busca estabelecer as relações entre língua e sociedade, fenômenos linguísticos variáveis podem funcionar como uma espécie de índice de pertencimento linguístico a uma dada comunidade. Labov (2008 [1972]) toma como objeto de estudo a comunidade de fala, que não é um grupo de falantes que utiliza as mesmas formas linguísticas, mas um grupo que compartilha os mesmos valores associados aos usos da língua, o que pode ser observado, por exemplo, pelos julgamentos de valor (positivo ou negativo) atribuídos conscientemente pelos falantes aos usos linguísticos. Por outro lado, Eckert (2000) propõe o estudo da variação centrada nas comunidades de prática, nas quais os indivíduos, ao escolherem pertencer a esta ou àquela comunidade, compartilham repertórios de práticas, dentre os quais as práticas linguísticas. A observação de comunidades de práticas permite ao pesquisador identificar como as variantes linguísticas assumem significado social, possibilitando estabelecer relação mais direta entre língua e significado do que em um estudo baseado em uma comunidade de fala, que, dado o seu delineamento, não permite

[1]O projeto *Falares Sergipanos* integra um projeto maior, intitulado *Da expressividade da língua ao mal na literatura: base de pesquisas interinstitucionais do PPGL/UFS* (FREITAG et alii, 2012), financiado pelo edital CAPES/FAPITEC/SE 06/2012, que visa a consolidar o Programa de Pós-Graduação em Letras da Universidade Federal de Sergipe (PPGL/UFS) com o incremento de sua infraestrutura de pesquisa e qualificação docente, em parceria com o Programa de Pós-Graduação em Linguística da Universidade Federal de Santa Catarina (PPGLg/UFSC) e o Programa de Pós-Graduação em Literatura da Universidade Federal de Minas Gerais (Pós-Lit/UFMG).

http://dx.doi.org/10.5007/1984-8420.2013v14n2p156

controlar as relações estabelecidas entre os falantes e suas implicações na dinâmica linguística.

Em sua teoria da variação tida como prática social, Eckert (2000) olha para os falantes como sujeitos que, ao se inserirem em práticas sociais, constituem categorias sociais e constroem (e respondem a) o significado social da variação. Com isso, é inerente ao fenômeno de variação/mudança linguística o processo de constituição da identidade dos indivíduos, pois é nesse processo (que envolve também a constituição do gênero) que as variáveis linguísticas assumem valor social.

O estudo da variação linguística como prática social requer, além da realização de análise quantitativa, a observação dos falantes em comunidades de prática. Nesse modelo de análise, a entrevista sociolinguística mostra-se instrumento relevante não apenas para coletar dados de fala, mas também para proceder a um primeiro diagnóstico dos grupos ou comunidades formadas em torno de um empreendimento comum. As narrativas de experiência pessoal favorecidas nas entrevistas sociolinguísticas fornecem pistas sobre a relação em rede (social) dos indivíduos e sobre os grupos em que se constituem as *personae* ou identidades sociais (ECKERT, 2012) reconhecidas em uma localidade.

Nas comunidades de prática, a liderança, por exemplo, pode dar ao líder o poder de propor inovações, até mesmo linguísticas, já que o grupo de liderados o legitima e o segue, aderindo aos comportamentos por ele adotados. É também nas comunidades de prática que se pode observar, através de estudo etnográfico, como as relações entre uso da linguagem, estilo e construção de identidade se dão para cada indivíduo.

A metodologia de pesquisa da sociolinguística variacionista, a fim de desvelar relações sistemáticas entre a variação linguística e a dinâmica social, foi amplamente difundida e aprimorada, especialmente no cenário brasileiro, chegando a tal ponto que a abordagem tem focado cada vez mais a dimensão sistemática da mudança linguística, esvaindo-se os valores sociais associados à variação. Recentemente, uma nova onda de estudos tem se dedicado a retomar a questão do estilo na variação; Eckert (2012) propõe uma retomada do significado social da variação, naquilo que denomina de estudos de terceira onda da sociolinguística.[2] Para tanto, Eckert sugere a mudança de foco: de **comunidades de fala**

[2]Em artigo seminal apresentado em 2005 no *Annual Meeting of the Linguistic Society of America*, em Oakland, Estados Unidos, e publicado definitivamente em 2012, Penelope Eckert apresenta uma proposta de tipologização da Sociolinguística em ondas, entendidas como tendências de estudos, que não se suplantam, nem que são sucessivas temporalmente, mas que mostram a gradação entre o social e o linguístico da Sociolinguística. A primeira onda de estudos sociolinguísticos é caracterizada pelos estudos de mapeamento de tendências amplas em comunidades de fala, com evidências quantitativas que buscam um padrão regular de distribuição de variantes associado a um perfil sociodemográfico específico, como no estudo laboviano da estratificação do inglês na cidade de Nova Iorque. Os estudos de segunda onda são também de orientação quantitativa e em comunidades de fala, mas adotam uma metodologia de base etnográfica, com coletas de dados mais longas; buscam também identificar um padrão regular de distribuição das variantes, mas correlacionado a categorias sociodemográficas mais abstratas, como a questão da identidade com a região, gostar ou não da cidade etc., como os estudos labovianos do inglês afroamericano. As primeira e segunda ondas de estudos sociolinguísticos partem da premissa de que as variáveis linguísticas carregam status social

para **comunidades de práticas**. A comparação dos resultados entre o estudo de comunidades de fala e o de comunidades de práticas, como a aqui proposta, permite a detecção de padrões de emergência em comunidades de fala e a observação da atuação de valores sociopessoais em comunidades de práticas. A confluência de abordagens tem sido testada em novos bancos de dados.

O estudo em comunidades de fala, com tendências amplas, permite que os resultados sejam aprofundados, desde que se tome como referência estudos microetnográficos de comunidades de práticas. Por entendermos as ondas de Eckert (2012) não como suplementares, mas complementares, a constituição de novos bancos de dados não pode abrir mão da comparabilidade com os bancos de dados já constituídos. Nessa perspectiva, o banco de dados *Falares Sergipanos* segue duas linhas de coleta – a de comunidades de fala (estratificação homogeneizada) e a de comunidades de práticas (relações sociopessoais). A amostra de estratificação homogeneizada é a que segue o padrão estabelecido nos bancos de dados sociolinguísticos brasileiros, com a identificação de informantes com um perfil específico – e cada vez mais raro de se encontrar, ressalte-se –, que são aqueles nascidos e criados na comunidade onde vivem, filhos de pais com as mesmas qualidades, estratificados quanto às características sociodemográficas amplas, como sexo, faixa etária e nível de escolaridade. A amostra de relações sociopessoais não segue essa estratificação rígida; a seleção se dá a partir do foco de interesses – a prática – em questão.

Atendendo às diretrizes norteadoras de pesquisa envolvendo humanos, normatizada e regulamentada no Brasil pela Resolução 196/96, o projeto *Falares Sergipanos* foi submetido à apreciação do Comitê de Ética em Pesquisa – CEP da Universidade Federal de Sergipe, o qual está vinculado ao Sistema Nacional de Informações sobre Ética em Pesquisa – SISNEP, recebendo certificado de atendimento às diretrizes éticas de pesquisa de 0386.0.107.000-11.

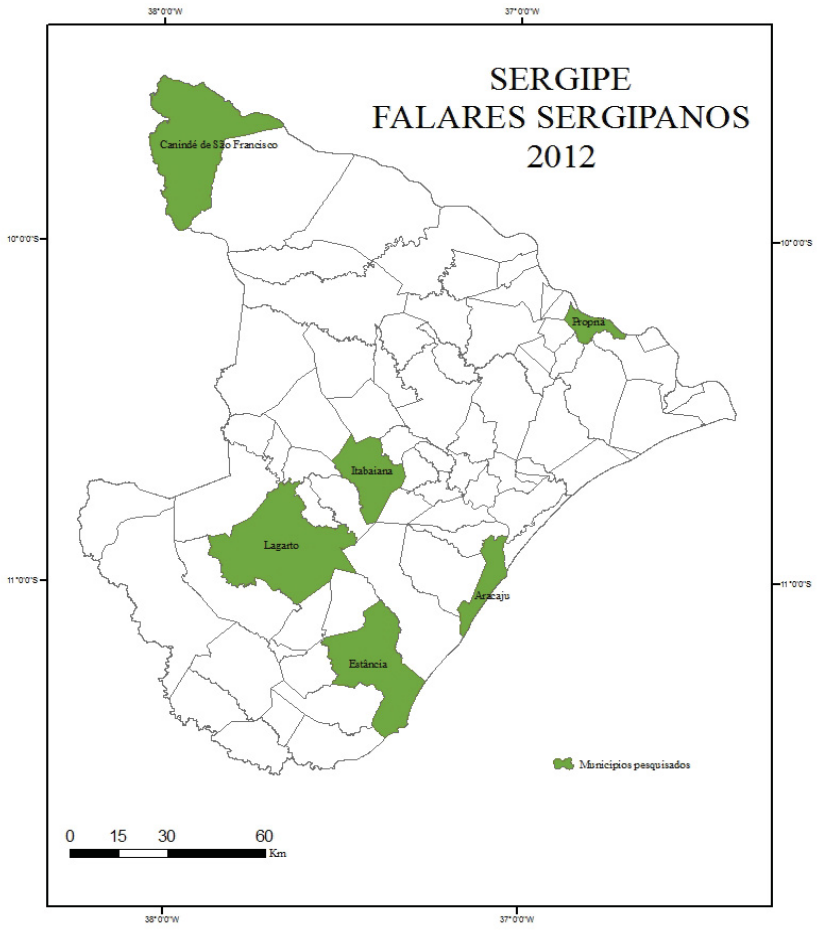
3.1 Abrangência da amostra

A coleta de comunidades de fala, nos moldes labovianos (LABOV, 2008 [1972]), com estratificação homogeneizada é predominante nos bancos de dados já constituídos; a sua replicação, portanto, viabiliza a comparabilidade de amostras de comunidades distintas. Para o dimensionamento da amostra, foram selecionadas seis cidades representativas do estado de Sergipe, por territórios: Canindé de São Francisco, Itabaiana, Lagarto, Estância, Propriá e a atual capital, Aracaju (Figura 1).

dos seus falantes; assim, o foco do estudo é a descrição da estrutura. Os estudos de terceira onda procuram investigar como os condicionamentos sociais impostos e as relações de poder atuam sobre as estruturas, a fim de identificar o significado social da variação, o estilo.

http://dx.doi.org/10.5007/1984-8420.2013v14n2p156

Figura 1 Distribuição das comunidades de fala



Fonte: Coleta de campo, 2012
Base de dados: SRH, 2011
Elaboração: Gicélia Mendes

3.2 Metodologia de coleta de dados e seleção de informantes

A estratificação etária dos informantes segue a padronização do Instituto Brasileiro de Geografia e Estatística – IBGE, computando cinco faixas (até 14 anos; 15-24; 25-39; 40-64; mais de 65 anos). A seleção dos informantes (inicialmente 2 para cada célula social) segue a abordagem "bola de neve" – em que um informante indica o outro informante, conforme o modelo de redes de Milroy e Milroy (1992), a partir do contato inicial do pesquisador de campo da comunidade. Não foi prevista, inicialmente, a estratificação por nível de escolarização, o que está sendo feito, ao ritmo da coleta, a partir de mapeamento qualitativo. Quando finalizado, o banco de dados contará com 40 entrevistas por cidade, totalizando 240 entrevistas sociolinguísticas, na amostra de comunidades de fala, e pelo mapeamento de 18 comunidades de práticas.

http://dx.doi.org/10.5007/1984-8420.2013v14n2p156

A metodologia de coleta prevê que os potenciais informantes passem por dois momentos: o da pré-seleção, com a realização de uma entrevista de checagem para o levantamento do perfil sociocultural do informante, e o da realização da entrevista sociolinguística propriamente, que segue um roteiro pré-estabelecido de sequências, com as perguntas de checagem (nome, idade, onde mora etc.), temas relacionados à infância, temas relacionados a risco de vida – de modo que remetam a tipos textuais de base narrativa –, temas relacionados a estudo, trabalho, política, cidade, entre outros – de modo que remetam a tipos textuais de base mais argumentativa/explanativa, além de perguntas sobre a comunidade (gosta/não gosta; quer ficar; o que pensa) e perguntas de avaliação da fala.

Realizar entrevistas sociolinguísticas não é uma tarefa fácil como se pode pensar a princípio, já que o objetivo do pesquisador ao realizar esse tipo de entrevista é coletar dados de fala o menos monitorada possível, ou seja, o vernáculo, momento em que o mínimo de atenção é prestado à língua, como postula Labov. Ao coletar tais dados, o pesquisador precisa lidar com dois empecilhos – a sua própria presença enquanto pesquisador e a presença do gravador –, gerando o que Labov chama de "paradoxo do observador", uma vez que a presença desses interfere na naturalidade da situação comunicativa. O treinamento dos entrevistadores tem sido uma alternativa que tem surtido efeitos para minimizar o paradoxo do observador.

A amostra de comunidades de práticas é constituída por grupos de indivíduos, nas cidades selecionadas para a constituição do banco de dados de comunidades de fala, em ação em comunidades de práticas em espaços específicos: i) em espaço escolar; ii) em espaço de trabalho; iii) em espaço recreativo; e iv) em espaço religioso. Tal amostra configura-se como dados de base interacional, distanciando-se do modelo laboviano. Os indivíduos são consultados sobre o seu consentimento de permissão de gravação, seguindo o protocolo do termo de consentimento livre e esclarecido, conforme orientações aprovadas para o projeto no Comitê de Ética em Pesquisa. Na constituição dessa amostra, foram realizadas gravações de longo termo, em intervalo semanal por um período de seis meses, a fim de captar nuanças de estilo e adequação de papéis sociopessoais dos participantes, com coletas longitudinais com os informantes.

3.3 Armazenamento, manipulação e disponibilização dos dados

Após constituídas as amostras, com a coleta de dados de campo, as entrevistas realizadas e as gravações das interações em comunidades de práticas foram submetidas aos procedimentos de validação, transcrição e revisão de transcrição. No procedimento de validação, a qualidade do áudio é aferida, a fim de identificar ruídos que porventura possam comprometer a amostra. Além disso, as informações sociais do informante são conferidas. O processo de transcrição é baseado na audição impressionística do áudio, mas, diferentemente de outros bancos de dados que seguem o protocolo de coleta da entrevista sociolinguística, que adotam uma transcrição de base fonética adaptada, no banco de dados *Falares Sergipanos* seguimos um modelo de transcrição tomando como referência os princípios ortográficos da escrita do português, o que nos permite fazer uso das ferramentas computacionais da Linguística de *Corpus* – como os *softwares* concordanceadores – e

tornar o trabalho de levantamento de dados mais otimizado. Os *corpora* da Sociolinguística que não adotam a transcrição ortográfica padrão dificultam o processamento das listas de frequência na Linguística de *Corpus*, limitando sua aplicação, assim como o fato de a informação social dos informantes não ser processável (MCENERY; HARDIE 2012; REPPEN; FITZMAURICE; BIBER, 2002). É possível buscar aproximações entre as áreas, com a padronização ortográfica de *corpora* sociolinguísticos e o desenvolvimento de etiquetas XML para codificar a informação social de modo a ser processável por *softwares* concordanceadores. Para possibilitar a manipulação dos dados, a transcrição ortográfica padrão é mais eficiente; marcações fonológicas podem ser feitas após a seleção dos contextos pela transcrição ortográfica, com tratamentos acústicos em *softwares* específicos, garantindo resultados mais acurados.

Para a sincronização do áudio e da transcrição, o pacote *Exmaralda*[3] foi a ferramenta que apresentou melhores resultados. Sob a chancela do Grupo de Estudos em Linguagem, Interação e Sociedade – GELINS, quando finalizado, o banco de dados *Falares Sergipanos* será disponibilizado à comunidade científica, como mais uma fonte para estudos descritivos de variedades do português falado.

CONSIDERAÇÕES FINAIS

Estudos sociolinguísticos que articulem resultados do comportamento de fenômenos variáveis em comunidades de práticas e em comunidades de fala têm sido objeto recente de investigação no cenário sociolinguístico brasileiro e se encontram ainda em estágio inicial. Ainda não existem bancos de dados constituídos sob essa perspectiva metodológica, o que torna a tarefa do banco de dados *Falares Sergipanos* ainda mais árdua e relevante, pois vai desenvolver protocolo piloto para balizar as próximas ações dentro dessa perspectiva. A modelagem da amostra, por contemplar comunidades de fala e comunidades de práticas, permite não só estudos de cunho quantitativo, mas também estudos de caráter mais etnográfico.

Além disso, é preciso destacar que a utilidade de bancos de dados sociolinguísticos vai além dos estudos descritivos da língua. Trata-se de uma fonte rica a ser explorada para: fins didáticos, proporcionando oportunidades de experiências com a heterogeneidade linguística do Brasil, à disposição do professor de Língua Portuguesa como língua materna e estrangeira; fins de estudo do processamento de linguagem natural, na medida em que o banco de dados *Falares Sergipanos* é constituído por textos autênticos e que podem ser manipulados por softwares concordanceadores.

[3]O pacote EXMARaLDA – EXtensible MARkup Language for Discourse Annotation – (SCHMIDT; WÖRNER, 2009) apresenta uma ferramenta denominada de Partitur, que, em forma de uma partitura, possibilita inserir, editar e produzir transcrições. A ferramenta Editor EXMARaLDA possibilita o intercâmbio de dados com outros *softwares*, como o Praat e o ELAN, para a segmentação dos dados de transcrição.

http://dx.doi.org/10.5007/1984-8420.2013v14n2p156

REFERÊNCIAS

BRASIL. Ministério da Educação e do Desporto. Secretaria de Educação Fundamental. *Parâmetros curriculares nacionais terceiro e quarto ciclos do ensino fundamental:* Introdução aos parâmetros curriculares nacionais. Brasília, DF: MEC/SEF, 1998.

ECKERT, P. *Linguistic variation as social practice*. Oxford: Blackwell, 2000.

_____. Three waves of variation study: the emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, n. 41, p. 87-100, 2012.

FREITAG, R. M. Ko. *et alii*. *Da expressividade da língua ao mal na literatura: bases de pesquisa institucionais do PPGL/UFS*. (Projeto submetido ao edital CAPES/FAPITEC/SE 06/2012). Universidade Federal de Sergipe, Programa de Pós-Graduação em Letras, 2012.

FREITAG, R. M. Ko; MARTINS, M. A.; TAVARES, M. A. Bancos de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: potencialidades e limitações. *Alfa*, n. 56, v. 6, p. 917-944, 2012.

GONÇALVES, S. C. L. Balanço crítico da Sociolinguística Variacionista no estado de São Paulo e a proposição de uma frente programática de investigação. *Estudos Linguísticos*, v.41, n.2, p. 869-884, 2012.

LABOV, W. *Padrões sociolinguísticos.* São Paulo: Parábola Editorial, 2008 [1972].

McENERY, T.; HARDIE, A. *Corpus linguistics*: method, theory and practice. Nova Iorque: Cambridge University Press, 2012.

MILROY, L.; MILROY, J. Social network and social class: toward an integrated sociolinguistic model. *Language in Society*, v. 21, n. 1, p. 1-26, 1992.

REPPEN, R., FITZMAURICE, S.; BIBER, D. *Using corpora to explore linguistic variation*. Amsterdã/Filadéfila: John Benjamins Publishing, 2002.

SALOMÃO, A. C. B. Variação e mudança linguística: panorama e perspectivas da sociolinguística variacionista no Brasil. *Fórum Linguístico*, v. 8, n. 2, p. 187-207, 2011.

SCHMIDT, T.; WÖRNER, K. EXMARaLDA – creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, v.19, n.4, p.565-582, 2009.

Recebido: 14/06/2013
Aceito: 31/08/2013