BANCOS DE DADOS SOCIOLINGUÍSTICOS DO PORTUGUÊS BRASILEIRO E OS ESTUDOS DE TERCEIRA ONDA: POTENCIALIDADES E LIMITAÇÕES

Raquel Meister Ko. FREITAG*
Marco Antonio MARTINS**
Maria Alice TAVARES***

- RESUMO: Bancos de dados linguísticos de fala – especialmente aqueles elaborados para a pesquisa de orientação sociolinguística variacionista – têm sido fonte privilegiada para a descrição do português brasileiro. Neste texto, discutimos procedimentos metodológicos que deveriam ser adotados para a organização de novos bancos de dados. Fazemos um breve retrospecto dos bancos de dados já constituídos e sugerimos a coleta e expansão de *corpora* de diferentes comunidades de fala – e de diferentes comunidades de prática. De acordo com proposta defendida por Eckert (2012), os estudos sociolinguísticos podem ser distinguidos em três ondas de análise que refletem modos distintos de abordagem à variação linguística. Sugerimos estratégias para padronizar os procedimentos de organização de bancos de dados sociolinguísticos que levem em conta as três diferentes ondas da pesquisa sociolinguística, e destacamos a terceira onda, ainda incipiente no Brasil. A padronização dos bancos de dados sociolinguísticos facilitaria a realização de investigações contrastivas de diferentes dialetos brasileiros, contribuindo, dessa forma, para o estabelecimento e refinamento de generalizações e princípios de variação e mudança universais.

- PALAVRAS-CHAVE: Sociolinguística. Banco de dados. Variação e mudança linguística. Fatores sociais. Estilo.

Introdução

Na literatura sociolinguística variacionista, a referência à metodologia costuma ocupar duas a três linhas, quando muito um parágrafo: "como *corpus* foram selecionados X informantes do banco de dados Y, estratificados em Z células sociais". A voz passiva da construção e a exiguidade do espaço dedicado à

* UFS – Universidade Federal de Sergipe, Centro de Educação e Ciências Humanas – Departamento de Letras Vernáculas. São Cristóvão – Sergipe – Brasil. 49100-000 - rkofreitag@uol.com.br

** UFRN – Universidade Federal do Rio Grande do Norte, Centro de Ciências Humanas, Letras e Artes. Natal – Rio Grande do Norte – Brasil. 59072-970 - marcoamartins.ufrn@gmail.com

*** UFRN – Universidade Federal do Rio Grande do Norte, Centro de Ciências Humanas, Letras e Artes. Natal – Rio Grande do Norte – Brasil. 59072-970 - aliceflp@hotmail.com

metodologia de organização do *corpus* não condizem com o real esforço e tempo dispendidos no processo de constituição de um banco de dados, desde a prospecção e seleção de informantes até a transcrição, armazenamento e disponibilização.

Bancos de dados linguísticos de fala (especialmente os que seguem a orientação da Sociolinguística Variacionista) têm sido fonte privilegiada para a descrição do português brasileiro. Porém, a constituição de *corpus* que procure considerar as variedades do português brasileiro é tarefa dispendiosa não só quanto a recursos financeiros, mas também quanto ao tempo. A cada projeto que constitui seu banco de dados em uma comunidade de fala, o mapeamento das variedades do português no Brasil vai se efetivando, mas só a padronização dos procedimentos metodológicos permitirá a realização de estudos contrastivos entre as variedades, para, então, possibilitar uma descrição mais acurada do português brasileiro.

Iniciativas para viabilizar estudos linguísticos variacionistas otimizando os recursos têm se tornado prática no cenário nacional, afinal, a comparação entre resultados obtidos para fenômenos variáveis é um recurso analítico que permite grandes avanços teóricos para a pesquisa linguística, uma vez que transcender os limites de uma única variedade linguística possibilita o estabelecimento, refinamento e fortalecimento de generalizações e princípios de variação e mudança universais (TAVARES, 2002).

Como um dos objetivos da Sociolinguística variacionista é obter resultados que possam ser generalizados, sua metodologia deve ser pautada em confiabilidade (os mesmos resultados devem ser repetidos na análise do mesmo fenômeno) e intersubjetividade (dois pesquisadores diferentes devem obter os mesmos resultados seguindo a mesma metodologia) (BAILEY; TILLERY, 2004). Ao tecerem avaliações de ordem metodológica – campo ainda pouco explorado nos estudos variacionistas – Bailey e Tillery discutem três razões possíveis para explicar a divergência de resultados em abordagens sociolinguísticas quantitativas a partir da premissa de que diferenças metodológicas resultam em divergências de resultados, especialmente quanto às diferenças de entrevistador e às diferenças na amostra da população.

Bailey e Tillery (2004) retomam o estudo de Rickford e McNair-Knox (1994), em que o mesmo informante afroamericano foi entrevistado por duas entrevistadoras: uma também afroamericana e outra branca. A frequência com que traços característicos do AAVE apareciam na entrevista realizada pela afroamericana era sensivelmente superior à da entrevista realizada pela entrevistada branca. Outro aspecto que Bailey e Tillery destacam é que a experiência do pesquisador de campo que realiza a entrevista (e, em menor efeito, o tópico) também mostra efeitos na frequência de fenômenos. Quanto aos efeitos da amostra, Bailey e Tillery (2004) ressaltam que os estudos sociolinguísticos costumam

trazer poucas informações acerca da seleção dos seus informantes ou de sua representatividade na comunidade. Os autores sugerem, a título de recomendação, que os pesquisadores precisam especificar exatamente qual é a amostra da população em estudo, assim como especificar quais os procedimentos para definir a amostra, de modo que possam garantir a confiabilidade e a intersubjetividade da análise. A discussão de Bailey e Tillery (2004) sugere que os resultados de uma investigação sociolinguística são não raro muito mais a consequência de escolhas metodológicas do que o comportamento dos informantes. O que torna a situação problemática, segundo os autores, é que a sociolinguística quantitativa não tem um corpo de pesquisadores que se dediquem ao método, nem literatura que explore os efeitos de diferentes entrevistadores, diferentes estratégias de elicitação dos dados, procedimentos de amostragem ou estratégias analíticas. À esteira da constatação de Bailey e Tillery (2004), neste texto, tecemos reflexões acerca da metodologia das abordagens sociolinguísticas de orientação variacionista desenvolvidas no Brasil, especificamente no que tange à constituição de bancos de dados. Fazemos um breve retrospecto dos bancos de dados já constituídos e prospectamos ações futuras neste campo, com a expansão e ampliação de amostras de variedades linguísticas (novos bancos de dados e coletas piloto). A orientação da discussão segue a proposta de Eckert (2012) a respeito das três ondas da Sociolinguística e as ponderações de Bailey e Tillery (2004), já apresentadas, sobre dados divergentes em sociolinguística (e que procedimentos metodológicos são pertinentes para minimizá-los).

As três ondas da Sociolinguística e os bancos de dados brasileiros

Propondo uma discussão sobre os rumos do significado social no estudo da variação, Eckert (2012)[1] faz uma abordagem programática dos estudos sociolinguísticos com o propósito de relevar o estudo da variação com ênfase no significado social: como o sistema de significado social é estruturado? Que tipos de significados sociais são expressos na variação? Em seu retrospecto, Eckert destaca que os estudos sociolinguísticos podem ser agrupados em três ondas de estudos, não substitutivas nem sucessivas, mas que se configuram como modos distintos de pensar a variação, com práticas analíticas e metodológicas peculiares. A proposta das três ondas dos estudos sociolinguísticos de Eckert vem recebendo sinalizações de que merece atenção no cenário sociolinguístico brasileiro (BENTES, 2009; CAMACHO, 2010; SCHERRE, 2011; HORA; WETZELS,

[1] Esse texto é a versão revisada e modificada do trabalho intitulado Variation, convention, and social meaning, que foi apresentado por Penelope Eckert na Annual Meeting of the Linguistic Society of America, em Oakland, no ano de 2005. Nessa versão mais recente de sua proposta, a autora sugere a integração entre os níveis social e cognitivo na dinâmica da variação que potencialmente os estudos de terceira onda poderiam abarcar. A versão mais antiga do texto está disponível em <http://www.stanford.edu/~eckert/EckertLSA2005.pdf>.

2011, entre outros), motivando a discussão acerca de seu impacto para o campo de estudos. Apresentamos, a seguir, a proposta de Eckert (2012) para a abordagem de cada uma das ondas de estudos sociolinguísticos, com ênfase nos aspectos metodológicos, especialmente no que diz respeito à constituição de bancos de dados sociolinguísticos brasileiros. Ao final da seção, avaliamos as potencialidades e as limitações de cada uma das abordagens quanto à constituição de bancos de dados, delimitando os aspectos que os novos bancos de dados potencialmente precisam contemplar para contribuir de modo efetivo aos estudos que enfocam o significado social da variação.

A primeira onda de estudos sociolinguísticos inicia com os estudos de Labov sobre a estratificação do inglês na cidade de Nova Iorque, cujos resultados foram replicados em uma série de estudos em comunidades urbanas que corroboraram um padrão regular de estratificação socioeconômica das variáveis, em que o uso das variantes não padrão está inversamente relacionado ao *status* socioeconômico do falante (ECKERT, 2012). A primeira onda estabeleceu uma base sólida para o estudo da variação, evidenciando as correlações entre variáveis linguísticas e categorias sociais primárias, como classe socioeconômica, sexo, idade, escolaridade etc. Os padrões regulares e sistemáticos de covariação social e linguística levantaram questões sobre relações sociais subjacentes às categorias sociais primárias, o que conduziu ao surgimento da segunda onda, caracterizado por estudos etnográficos de populações mais localmente definidas.

A premissa dos estudos de primeira onda é, pois, que as variedades linguísticas carregam o status social de seus falantes. A metodologia dos estudos de primeira onda é calcada na correlação entre as variáveis linguísticas e as categorias socioeconômicas em sentido amplo (cuja classificação se dá de forma estável, homogênea e padronizada de modo a permitir a replicação, como faixa etária, sexo, etnicidade, escolaridade), com a estratificação dos falantes em células sociais, a constituição de bancos de dados linguísticos e resultados quantitativos refinados (especialmente com o uso de técnicas estatísticas aprimoradas para o modelo da variação linguística, como a regressão logística com o cálculo de desvio da média ponderada (SANKOFF; TAGLIAMONTE; SMITH, 2005)).

No Brasil, os estudos quantitativos com bancos de dados estratificados de acordo com características sociodemográficas amplas têm se consolidado como modelo hegemônico, com os bancos de dados do Programa de Estudos sobre o Uso da Língua (PEUL), da Universidade Federal do Rio de Janeiro, que foi o pioneiro no Brasil a implementar esse modelo de constituição de amostra[2]. A partir deste, foram replicados projetos em diferentes regiões do Brasil, com adaptações em sua metodologia (Projeto Variação Linguística Urbana na Região

[2] Para mais detalhes sobre o banco de dados PEUL, ver Scherre e Roncarati (2008).

Sul do Brasil – VARSUL, da equipe formada pela Universidade Federal do Paraná, Universidade Federal de Santa Catarina; Universidade Federal do Rio Grande do Sul e Pontifícia Universidade Católica do Rio Grande do Sul[3]; Projeto Variação Linguística na Paraíba – VALPB, da Universidade Federal da Paraíba; Banco de Dados Sociolinguístico da Fronteira e da Campanha Sul-Rio-Grandense – BDS-Pampa, da equipe da Universidade Federal de Pelotas e da Pontifícia Universidade de Pelotas; Banco de Dados por Classe Social – VarX, da Universidade Federal de Pelotas, entre outros).[4] Esse tipo de banco de dados possibilita captar tendências amplas de variação e mudança em uma comunidade de fala. Implica, entretanto, a homogeneização da amostra, como discutimos mais à frente.

É importante destacar que a elaboração desses *corpora* permitiu a descrição do português brasileiro em diferentes aspectos linguísticos e considerando distintas variedades. De algum modo, têm-se uma descrição da variação na(s) gramática(s) do português do Brasil envolvendo diferentes fenômenos e a correlação destes com variáveis sociais.

Os estudos de segunda onda são também de natureza quantitativa, mas de abordagem etnográfica, abarcando categorias sociodemográficas mais abstratas, a fim de evidenciar como o vernáculo assume valor local. Os estudos etnográficos enfocam comunidades menores por períodos de tempo relativamente longos com o objetivo de descobrir as categorias sociais localmente mais salientes. Essas categorias podem ser instanciações locais das categorias primárias que guiam os estudos quantitativos, mas o traço distintivo crucial desse tipo de estudo é a descoberta do lugar dessas categorias na prática social local. Nesse tipo de abordagem, o foco recai nos conceitos de comunidades de fala e de identidade de grupo. Eckert (2012) traz, em referência à segunda onda, três exemplos: (1) o estudo de Labov sobre o inglês afroamericano (AAVE), cujos resultados apontam para o uso de traços vernaculares por adolescentes como indexadores do status entre o grupo de comunidade de prática; (2) o estudo de Milroy (1980), que enfoca comunidades de classe operária e examina a relação entre engajamento local e uso do vernáculo, correlacionando o uso de variáveis vernaculares locais com a densidade e a multiplicidade da rede de relações sociais do falante; e (3) o estudo da própria Eckert sobre o papel das categorias *jokers* e *burnouts* na indexação de classe socioeconômica em grupos adolescentes (ECKERT, 2000).

No cenário brasileiro, esse tipo de abordagem não recebeu a mesma ênfase que os estudos quantitativos baseados em categorias sociais amplas. Dentre os poucos estudos que se encaixam na segunda onda, escolhemos o de Ferrari (1994) para ilustração. Ferrari (1994) selecionou doze traços fonológicos, escalonados entre

[3] Ver Bisol, Menon e Tasca (2008).

[4] Nesta relação, não incluímos o projeto Norma Urbana Culta – NURC porque, apesar de subsidiar também descrições de cunho sociolinguístico, seu banco de dados não foi constituído para essa finalidade.

discretos – em que a variável indica uma delimitação nítida entre grupos sociais contíguos – ou gradientes – em que a variável não se apresenta com frequência significativamente maior de um grupo social para outro –, e um traço sintático-semântico (variação de preposição locativa "em" vs. "ni"), a fim de verificar as relações entre variação e redes sociais na comunidade do Morro dos Caboclos, no Rio de Janeiro. A rede de relações sociais do indivíduo estabelecida na comunidade não se configura como um indicador sociodemográfico amplo, como sexo, idade, escolarização etc.; trata-se de um indicador que só é captado com um estudo investigativo individualizado, aos moldes etnográficos. Os resultados da investigação de Ferrari (1994) apontam que redes sociais relativamente fechadas possibilitam a focalização de traços linguísticos (conservação dos traços linguísticos característicos da comunidade do Morro dos Caboclos), enquanto redes sociais pouco coesas associam-se à difusão linguística (abandono dos traços linguísticos da comunidade em troca de traços que os aproximam dos moradores de bairros da zona oeste carioca, nas proximidades do Morro dos Caboclos): os falantes que trabalhavam na cidade faziam uso de traços linguísticos diferentes daqueles que nunca desciam o morro. Estudos dessa natureza permitem uma avaliação mais acurada do fenômeno da variação, com ênfase no valor social das variáveis. São, entretanto, estudos dispendiosos e demorados, cujo *corpus* de análise não segue o alinhamento dos bancos de dados constituídos de acordo com a estratificação social baseada em indicadores sociodemográficos amplos.

Os estudos de primeira e segunda ondas, segundo Eckert (2012), têm como foco a descrição da estrutura – um retrato estático. Os estudos de terceira onda incorporam a dinamicidade da estrutura, ou seja, como a estrutura se molda no cotidiano, com os condicionamentos sociais impostos e as relações de poder estabelecidas atuando sobre ela. Eckert (2012) salienta que não está negando a estrutura, mas sim enfatizando o papel da estrutura no condicionamento da prática paralelamente ao papel da prática na produção e reprodução da estrutura, a fim de captar com mais detalhes a dinâmica do valor social das variáveis.

Os estudos de terceira onda combinam os postulados dos estudos de primeira e de segunda onda, com uma mudança no foco: da comunidade de fala para a comunidade de prática. Enquanto, na definição laboviana, comunidades de fala são agrupamentos de indivíduos que compartilham não necessariamente dos mesmos traços linguísticos, mas sim do mesmo juízo de valor acerca desses traços, e os reconhecem como legítimos para a identificação do grupo, a comunidade de prática (WENGER, 1998; ECKERT; MCCONNELL-GINET, 2010; ECKERT; MCCONNELL-GINET, 1997) é um agrupamento de indivíduos (comunidade) que partilham perspectivas em comum, valores e conhecimento (domínio), e que interagem entre si para se aperfeiçoarem e replicarem esses valores e conhecimentos (prática). Trata-se de uma construção social, e, como tal, está sujeita às práticas diárias dos indivíduos, que interagem entre si e com outras comunidades.

> Em lugar de conceber o indivíduo como uma entidade à parte, pairando sobre o espaço social, ou como um ponto em uma rede, ou como membro de um conjunto específico ou de um conjunto de grupos, ou como um amontoado de características sociais, precisamos enfocar as comunidades de prática. Tal foco possibilita-nos ver o indivíduo como agente articulador de uma variedade de formas de participação em múltiplas comunidades de prática. (ECKERT; MCCONNEL-GINET, 2010, p.103).

A terceira onda, que se desenvolveu mais recentemente, centra o foco na variação vista não como o reflexo do lugar social num ponto da escala, mas como um recurso para a construção de significado social. Eckert (2012) se volta à necessidade de conectar essas categorias sociais mais abstratas, arraigadas na experiência do falante, com as comunidades imaginárias mais amplas, centrando foco na construção do conceito de comunidade de prática. Uma comunidade de prática é um agregado de pessoas que se juntam para engajar-se em algum empreendimento comum. Na esteira desse engajamento, a comunidade de prática desenvolve meios para fazer coisas que se traduzem em práticas e essas práticas envolvem a construção de uma orientação compartilhada em relação ao mundo em volta – uma definição tácita que os indivíduos assumem um em relação ao outro e em relação a outras comunidades de prática.

Os estudos de terceira onda combinam a metodologia quantitativa, presente nas ondas anteriores, o *corpora* constituídos de modo a contemplar a dimensão mais cotidiana (o que não é necessariamente captado pela entrevista sociolinguística), com observações participantes, por exemplo.

O conceito-chave para o processo de construção é o de prática estilística. Até aqui, nos estudos variacionistas, o estilo tem sido tratado como ajustes à (in) formalidade da situação mediante o uso de variáveis individuais. A face renovada de estilo o identifica com o modo como os falantes combinam variáveis para criar modos distintivos de fala, que fornecem a chave para a construção da identidade. A identidade consiste, por sua vez, em tipos particulares explicitamente localizados na ordem social. Continuamente, os falantes atribuem significado social à variação de um modo consequente, situação que implica certo grau de agentividade.

Eckert (2012) postula que toda variação tem potencial para receber significado social, ainda que nem toda variação seja conscientemente controlada ou mesmo socialmente significativa. A indexação de variáveis fonológicas não é tão transparente quanto, por exemplo, o uso de partículas honoríficas, mas é justamente a fluidez das primeiras que as torna acessíveis a uma grande variedade de propósitos sociais. É necessário haver apenas tempo e continuidade suficientes para convencionar a relação entre uma variável e um significado social. É por essa razão que variáveis estáveis, como a redução de (*-ing*) no inglês americano, têm

significados tão extremamente claros que podem ser referidos como estereótipos, ao passo que variantes representando mudanças em progresso são recursos mais instáveis, mais transitórios e, por isso, mais disponíveis para assumirem significado social. Um contínuo da convencionalização acompanha um contínuo de intencionalidade, num processo que torna o sujeito agente dos processos sociais que constroem sua própria identidade.

Na linha dos estudos de terceira onda, Moore (2010) analisou a variação entre *were/was* em uma comunidade de prática em Bolton, Inglaterra. Seus resultados globais seguem o padrão da variação *were/was* obtidos por estudos baseados em bancos de dados sociolinguísticos (TAGLIAMONTE, 1998; CHESHIRE; FOX, 2009). Porém, sua metodologia de coleta etnográfica possibilitou captar a correlação entre o uso não padrão e a estrutura social da comunidade de prática, configurando a variável como um índice de prática social.

A investigação de Moore (2010) intenta mostrar como os fatores sociais interagem e avaliar como cada fator condiciona o uso de *were* em contextos de primeira e terceira pessoa do singular. A constituição da amostra se deu em um período de dois anos de observação etnográfica de adolescentes da *Midlan High School*, em Bolton, Inglaterra. Especificamente, foram consideradas as gravações de fala de 39 garotas. Para coletar esses dados, primeiramente a pesquisadora foi à escola no horário do almoço e se envolveu em atividades diversas (como almoçar na cantina, assistir a um ensaio de peça de teatro, sair com os fumantes para a área externa). Nessa etapa foram tomadas notas de campo; interações com as adolescentes só foram gravadas após seis meses de observação etnográfica, e nunca foram realizadas sob a forma de uma entrevista sociolinguística clássica, mas sob a forma de um grupo de discussão e relato de atividades, envolvendo entre duas e quatro adolescentes, o que resultou em 50 horas de gravação. Além disso, questionários circularam entre as participantes do estudo, para coletar informações sobre as práticas sociais das informantes (a fim de validar as observações etnográficas), além de informações sobre classe social, identificação de si mesmo, de família e da sua localidade de nascimento.

A partir dessa observação etnográfica, Moore (2010) identificou quatro comunidades de prática, e cada adolescente membro foi avaliada quanto a com que ocupava seu tempo na escola, com quais atividades ela se engajava, sua orientação em relação aos pares e entorno, seu estilo pessoal e sua aparência e sua avaliação (por si mesma e pelos membros do grupo).

As "populares" exibem uma atitude antiescola, têm um estilo esportivo e feminino de se vestir e se engajam em atividades moderadamente rebeldes, como beber e fumar. Moore (2010) destaca que, no meio do trabalho de campo, as "*townies*" fundiram-se com as "populares", quando estas começaram a se engajar em atividades de certo risco, como envolvimento com drogas e atividade sexual.

Os amigos das "*townies*" são rapazes mais velhos, com quem elas ocupam seu tempo na escola. As "*geeks*" exibem uma atitude positiva em relação à escola, engajando-se em atividades como a banda da escola, esportes etc. Por fim, as "*Eden Village*", assim nomeadas em função de residirem em bairro de status, também são orientadas para valores institucionais da escola, e se vestem de acordo com o estilo *teen* da moda, ocupando seu tempo com dança, compras e festas de pijama.

Após analisar os dados coletados considerando fatores linguísticos e sociais, controlando subamostras e especialmente como cada adolescente se identifica e é identificada pelo grupo, Moore (2010) traz evidências quantitativas (baseada na frequência de uso individual) de que a comunidade de prática "*townie*" tende a fazer uso da forma não padrão (*were*), a comunidade de prática das "populares" se mostra neutra, enquanto a comunidade "*geeks*" assume uma postura desfavorável à forma não padrão e a comunidade "*Eden Village*" tem um padrão de uso da forma padrão (*was*) muito próximo do categórico.

Moore (2010) conclui que, para analisar o fenômeno de modo mais eficiente, é preciso observar o que se passa nas relações além do contexto institucional. No caso da variação entre *was* e *were*, os resultados globais se aproximam ao que outros estudos constataram; a distribuição considerando as comunidades de prática permite observar que a identidade não é apenas uma entidade social que é correlacionada a aspectos linguísticos, mas um fenômeno sociolinguístico, que é construído com o valor simbólico de características sociais e linguísticas.

Estudos de terceira onda têm tomado como objeto comunidades de prática variadas, como *yuppies* em Beijin (ZHANG, 2008), *gays* (PODESVA, 2002); PODESVA; ROBERTS; CAMPBELL-KIBLER, 2002), adolescentes em escolas (ECKERT, 2000; MOORE, 2010), que não podem ser aprioristicamente definidas, como vimos na descrição detalhada do estudo de Moore (2010).

O cenário sociolinguístico brasileiro atual vem acenando com entusiasmo para os estudos de terceira onda. Camacho (2010, p.160) diz que "O entendimento que temos da teoria sociolinguística permite assumir que o terceiro ciclo, na visão de Eckert (2005), é o ponto de vista mais consistente com o postulado de que a linguagem é um sistema adaptativo." É possível que o entendimento de Camacho (2010) seja acertado – apesar de não ser baseado em resultados de estudos empíricos no português –; porém, é preciso reconhecer que os estudos baseados em categorias sociodemográficas amplas (os de primeira onda) são particularmente importantes para respaldar os estudos de terceira onda. Apenas para exemplificação dessa importância, vejamos o estudo de Bentes (2009, p.118-119):

> Esta análise, que considera necessária a articulação entre diferentes recursos e níveis de linguagem para a explicação de elaboração de

> registros e de estilos linguísticos (sejam eles cultos ou populares), insere-se na agenda de estudos sociolinguísticos da chamada "terceira onda" (ECKERT, 2005), que pretende dar visibilidade aos complexos processos de elaboração de identidades, registros e estilos a partir da manipulação dos recursos das diferenças linguísticas no interior dos grupos sociais. (COUPLAND, 2001; BELL, 2001).

Ao analisar a fala de Mano Brown, Bentes (2009) foca traços linguísticos que associa ao português não padrão, como ausência de concordância explícita de número, e conclui que o comportamento do sujeito quanto ao traço considerado "apenas corrobora a tendência já afirmada em estudos sociolingüísticos" (NARO; SCHERRE, 2007; SCHERRE; NARO, 2007 apud BENTES, 2009, p.126). Os estudos referidos pela autora tomam por *corpus* amostras de bancos de dados sociolinguísticos – especificamente o PEUL –, o que os caracteriza como estudos de primeira onda. Hora e Wetzels (2011) também destacam a importância dos estudos de primeira onda para os estudos de terceira: ao analisarem os efeitos estilísticos da variação – em uma abordagem que procura se alinhar aos estudos de terceira onda – entre o uso de oclusivas dentais e africadas na fala de João Pessoa, os autores dizem que "Neste estudo, os dados coletados na Paraíba (VALPB) refletem esse momento que Eckert denomina de primeira onda." (HORA; WETZELS, 2011, p.162). Assim, apesar da visibilidade e da ênfase aos estudos alinhados à tendência da terceira onda dos estudos sociolinguísticos, os estudos de primeira onda – e, particularmente, as abordagens baseadas em bancos de dados sociolinguísticos – continuam a ter um significativo papel e importância na sociolinguística brasileira.

De um modo geral, bancos de dados constituídos de acordo com a metodologia da Sociolinguística são, ainda, importantes fontes para os estudos sociolinguísticos. Faz-se necessário, no entanto, aprimorá-los para contemplar a dimensão da comunidade de prática, do estilo e da *personae*, ou da terceira onda, nos termos de Eckert. Reflexões nessa direção são tecidas na seção a seguir.

Bancos de dados sociolinguísticos no Brasil: potencialidades e limitações

Como vimos, apesar de a dinâmica dos estudos de terceira onda focar relações entre estrutura e prática, a pesquisa sociolinguística baseada em bancos de dados segue mantendo seu espaço no cenário brasileiro, na medida em que possibilita captar tendências amplas em uma comunidade de fala. Essa relação implica, entretanto, a homogeneização da amostra e suas consequências supergeneralizantes.

A discussão sobre a homogeneização da amostra de comunidades de fala não é recente. Severo (2009, p.16) faz um retrospecto, apontando que atualmente

"[...] as pesquisas sociolinguísticas de variação/mudança (incluindo as labovianas) têm valorizado as dimensões micro de estudo, sendo que as unidades de análise deixam de se centrar na comunidade de fala, para integrar as ideias de redes sociais e de comunidades de prática."

Bancos de dados baseados em comunidades de fala caracterizam-se pela seleção aleatória de seus informantes, que sejam nascidos na comunidade e onde tenham vivido pelo menos 2/3 da vida; que sejam filhos de pais com as mesmas características, além de serem reconhecidos pelos pares como membro da comunidade de fala. A estratificação dos informantes se dá em função de características sociodemográficas (sexo, idade, escolaridade etc.), gerando células sociais (confluência de fatores estratificadores), que devem tender à ortogonalidade (GUY, 2007). Uma amostra estatisticamente representativa da comunidade de fala precisa contar com 0,5% do total da população, margem de erro assumida nas ciências humanas.

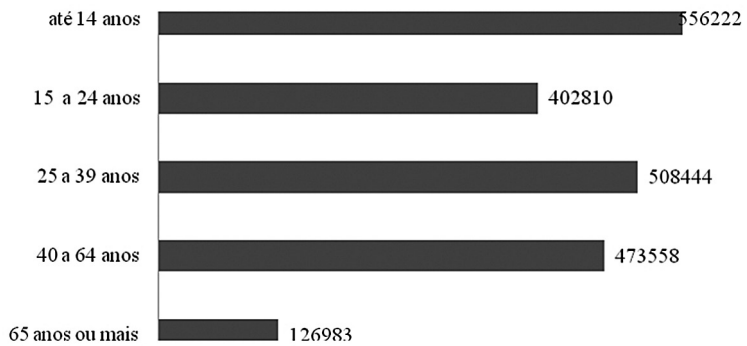
Freitag (2011b) também discute aspectos acerca da metodologia da constituição de amostras para bancos de dados sociolinguísticos:

> Se os bancos de dados têm como objetivo subsidiar a descrição de uma dada variedade de língua, e esta descrição, por conta da orientação teórico-metodológica, contempla a dimensão social, será que a estratificação das amostras homogeneizadas, como nos bancos de dados do PEUL e do VARSUL, reflete a estrutura social do Brasil? (FREITAG, 2011b, p.44).

Apenas a título de ilustração, vejamos a aplicação dos critérios de amostragem e representatividade assumidos pelos bancos de dados na aplicação aos bancos de dados *Falantes Universitários de Itabaiana/SE* e *Falares Sergipanos*, que serão apresentados com mais detalhamento na seção a seguir. Suponhamos que um banco de dados seja constituído a partir dos seguintes indicadores sociodemográficos: sexo, idade, escolaridade e zoneamento. Esse perfil social (dois sexos, três escolaridades, cinco faixas etárias e dois zoneamentos) geraria 60 células sociais. Idealmente, são necessários cinco falantes por célula social para garantir a confiabilidade e a representatividade da amostra (MOLLICA; BRAGA, 2004); a condição mínima de constituição de células sociais prevê dois falantes por célula (este é o padrão adotado no banco de dados do projeto VARSUL; o banco de dados do PEUL apresenta amostras com três falantes). Assim, na condição metodológica ideal, a amostragem seria constituída por 300 falantes; na condição mínima, 120 falantes. Agora vejamos o quão fidedigna é a amostragem constituída: tomando por base os dados do Censo 2010, realizado pelo Instituto Brasileiro de Geografia e Estatística – IBGE (2010), publicados no Diário Oficial da União do dia 04/11/2010 e disponibilizados na internet para consulta interativa – , a população

do município de Itabaiana/SE é de 86.019 habitantes; aplicando-se o corte de 0,5% (o mesmo que Labov usou em Martha's Vineyard), a amostra representativa da comunidade de fala deveria ser constituída por 430 falantes. A população do município de Aracaju/SE é de 552.365 habitantes; aplicando-se o corte de 0,5%, a amostra representativa da comunidade de fala deveria ser constituída por 2.762 falantes. A amostra ideal para Itabaiana/SE está relativamente próxima da amostra estatisticamente significativa; já para Aracaju/SE, a diferença é sensivelmente alta. O quão fidedigna é uma modelagem que homogeneiza duas amostras de populações sensivelmente diferentes? As distorções da homogeneização vão além: vejamos a distribuição da população do estado de Sergipe quanto à estratificação etária, considerando os dados do Censo 2010, apresentados no gráfico 1.

Gráfico 1 - Distribuição da população do estado de Sergipe por faixas etárias.



Fonte: Elaboração própria com dados do IBGE (2010).

A distribuição da população por faixas etárias apresentada no gráfico 1 aponta para uma redução sensível da faixa etária mais velha (com 65 anos ou mais), com aumento de população na faixa etária de até 14 anos. Se a amostragem do banco de dados respeitasse a proporção da estratificação da população, em um cenário de quatro indivíduos por célula social, a faixa etária de 65 anos ou mais teria apenas um representante, e a faixa etária até 14 anos teria cinco representantes. A opção metodológica por homogeneizar as amostras tem por implicação a restrição, ou ressalva, da generalização dos resultados, embora reconheçamos trabalhos que assumam a supergeneralização. As ponderações acima não significam a condenação dos bancos de dados constituídos nessa perspectiva; ao contrário, são registros sistemáticos e altamente produtivos para a identificação de tendências na comunidade, motivo que reforça a necessidade de continuidade desse trabalho de armazenagem, mas não sem uma readequação metodológica de

modo a contemplar as premissas da terceira onda. Na seção a seguir, discutimos aspectos metodológicos que devem ser considerados na constituição de novos bancos de dados sociolinguísticos (e também na ampliação dos já existentes), de modo que seja possível a comparabilidade entre as amostras.

Como vimos destacando, bancos de dados linguísticos têm sido fonte privilegiada para a descrição do português brasileiro e a tendência recente dos estudos de terceira onda ratifica sua importância para apontarem tendências linguísticas na comunidade. Seguindo as premissas da confiabilidade e da intersubjetividade (BAILEY; TILLERY, 2004), para dar continuidade a essa prática produtiva, novos bancos de dados precisam conservar minimamente as estratificações dos bancos de dados já existentes, pois a comparação de dados em tempo real permite análises mais acuradas com estudos de painel e de tendência (LABOV, 2001). Assim, as variáveis demográficas amplas – sexo, idade, escolarização etc. – precisam continuar a ser controladas nas novas coletas e nas novas amostras constituídas; é desejável, entretanto, que as novas coletas aprimorem o controle do falante, suas características individuais e de práticas, de modo a permitir que se construa um perfil social que contemple indicadores sociodemográficos mais amplos e abstratos. Um exemplo desse tipo de controle é o banco VarX: "[...] a construção do VarX surgiu da necessidade de se estudar com mais profundidade aspectos referentes a classes sociais (ocupação/profissão, renda/patrimônio e escolaridade) e suas implicações linguísticas." (AMARAL, 2003, p.63). Nesse banco de dados, informações mais acuradas sobre a região onde o falante mora, se o tipo de sua casa corresponde ao padrão do zoneamento, que tipos de bem de consumo possui, se sua profissão é manual, técnica ou intelectual, correlacionadas ao nível de escolarização, possibilitam um enquadramento mais próximo da realidade social do informante, desfazendo a homogeneização da amostra. O banco de dados VarX estratifica esses indicadores, mas a simples inclusão de mais questões relacionadas a fatores subjetivos nos questionários/roteiros de entrevistas sociolinguísticas já torna possível que esse tipo de informação seja extraída das amostras. É nessa linha também que vem seguindo o banco de dados Português Paulistano (MENDES, 2011), cujo objetivo é a constituição de um *corpus* contemporâneo do português paulistano que permita a sua descrição e análise nos moldes da sociolinguística variacionista.

Os novos bancos de dados sociolinguísticos têm também investido em coletas que privilegiem a diversidade de tipos/sequências textuais, de modo a captar estilos linguísticos mais próximos do dia a dia; é o caso do projeto Amostra Linguística do Interior Paulista – ALIP, cujo banco de dados Iboruna é constituído por amostras do português falado na região de São José do Rio Preto, e cidades circunvizinhas, na região noroeste do Estado de São Paulo.

Em relação à tendência de homogeneização da amostra, é importante destacar que o Iboruna é um banco de dados em que a distribuição dos informantes por célula social se dá proporcionalmente à densidade populacional das cidades da região, como podemos ver no quadro 1.

Quadro 1 - Distribuição dos informantes do bando de dados IBORUNA.

Cidades da Região de São José do Rio Preto	População	Número de informantes
1. Bady Bassit (12 km ao sul de SJRP)	11.475	04
2. Cedral (14 km, ao sul de SJRP)	6.690	02
3. Guapiaçu (16 km, ao leste de SJRP)	14.049	05
4. Ipiguá (18 km, ao norte de SJRP)	3.461	01
5. Mirassol (14 km, a oeste de SJRP)	48.233	16
6. Onda Verde (25 km, ao norte de SJRP)	5.407	02
7. São José do Rio Preto	357.705	122
Total da população representada	447.020	152

Fonte: Gonçalves (2008, p.2729).

O cuidado em dimensionar a amostra de modo a garantir uma representação proporcional da população, assim como as informações do perfil dos falantes quanto a fatores sociodemográficos mais amplos colaboram para a confiabilidade e a intersubjetividade da análise, de acordo com as recomendações de Bailey e Tillery (2004), apresentadas na introdução deste texto.

Para respaldar estudos de terceira onda, novos bancos de dados precisam também promover a mudança do foco da comunidade de fala para a comunidade de prática. Eckert e McConnel-Ginet (2010) explicitam o que entendem por comunidades de prática:

> Uma comunidade de prática pode ser constituída por pessoas trabalhando juntos em uma fábrica, *habitués* de um bar, companheiros de brincadeira em uma vizinhança, a família nuclear, parceiros policiais e seu etnógrafo, a Suprema Corte etc. Comunidades de prática podem ser grandes ou pequenas, intensas ou difusas; elas nascem e morrem, podem sobreviver a muitas mudanças de membros e podem estar intimamente articuladas a outras comunidades. As pessoas participam de múltiplas comunidades de prática, e a identidade individual é baseada nesta participação. Em lugar de conceber o indivíduo como uma entidade à parte, pairando sobre o espaço social, ou como um ponto em uma rede, ou como membro de um conjunto específico ou de um conjunto de grupos, ou como um amontoado de características sociais, precisamos enfocar as comunidades de prática. Tal foco possibilita-nos ver o indivíduo como agente articulador de uma variedade de formas de participação em múltiplas comunidades de prática. (ECKERT; MCCONNEL-GINET, 2010, p.102-103).

A constituição de *corpus* de comunidade de prática permite, por exemplo, a depender do tamanho da comunidade, que sejam considerados todos os indivíduos. Permite, também, que se proceda ao mapeamento acurado das redes de relacionamento, observando os graus de integração dos indivíduos dentro da comunidade de fala.[5]

Apesar dos pontos favoráveis apresentados à abordagem de comunidades de prática, a sua implementação única e exclusiva como fonte para os estudos sociolinguísticos não é benéfica; além de quebrar a série histórica da comparabilidade de amostras de comunidades de fala (com coletas iniciadas na década de 1980, como vimos na seção anterior), a abordagem de comunidades de prática, sem um estudo anterior baseado em comunidades de fala para levantar a(s) tendência(s) ampla(s) a ser(em) analisada(s), é um tiro no escuro que pode ou não resultar em uma boa investigação sociolinguística. Cada uma das abordagens apresenta suas particularidades e suas especificidades, como sistematizamos no quadro 2.

Quadro 2 - Comparação entre abordagens sociolinguísticas de comunidades de fala e de comunidades de práticas.

Abordagem de comunidade de fala	Abordagem de comunidade de práticas
- estratificação baseada em fatores sociodemográficos amplos	- estratificação baseada em valores localmente estabelecidos
- distribuição homogênea, tanto quanto ao tamanho quanto às categorias controladas	- distribuição variável, definida caso a caso
- categorias definidas a priori	- categorias definidas a posteriori
- permissão para captar tendências amplas da comunidade	- permissão para captar valores sociais localmente estabelecidos nas relações
- coleta padronizada (entrevista sociolinguística)	- coleta etnográfica (observação participante, interações entre grupos)
- constituição da amostra em curto prazo	- constituição da amostra em longo prazo

Fonte: Elaboração própria.

Como podemos facilmente constatar, cada uma das abordagens apresenta peculiaridades que não permitem a implementação simultânea, por conta de incompatibilidades de natureza teórico-metodológica. Entendemos, no entanto,

[5] Milroy e Gordon (2003) diferenciam as abordagens de investigação de redes sociais das de comunidades de prática: enquanto, em redes sociais, o objetivo é identificar o nó social que é importante para o indivíduo; em comunidades de prática, o objetivo é identificar o agrupamento que forma o local da prática linguística e da prática social.

que é viável a articulação entre as abordagens, e que essa articulação deve constar como proposta programática nas novas coletas de dados. O ponto de partida é, no entanto, a abordagem de comunidades de fala, pois é a partir dessa coleta que é possível delinear as tendências amplas da comunidade (e garantir a comparabilidade entre amostras, na medida que é adotada uma metodologia já consolidada) e captar pistas para definir abordagens de comunidades de prática.

Nas seções a seguir, nos dedicamos à apresentação de três novos bancos de dados no nordeste brasileiro – Bancos de fala culta de Itabaiana/SE, Falares Sergipanos, e FALA-Natal –, contemplando aspectos metodológicos da terceira onda de estudos sociolinguísticos.

Banco de fala culta de Itabaiana/SE

Localizada na região do agreste central sergipano, a cidade de Itabaiana é a cidade mais importante do Estado de Sergipe fora da região da Grande Aracaju. Abrigando uma central de abastecimento e uma feira de porte significativo para a região, a cidade é conhecida por sua fama de "comércio forte", atuando como entreposto comercial na circunvizinhança. O itabaianense tem uma atitude muito positiva quanto a si e quanto à sua cidade (FREITAG; SANTOS; SANTOS, 2009). Apesar de já haver um polo universitário particular, a implantação do *campus* de Itabaiana da Universidade Federal de Sergipe, em 2006, decorrente do programa do governo federal de expansão e interiorização da educação superior no Brasil, provocou grandes alterações na cidade de Itabaiana e circunvizinhança. Com oferta anual em dez cursos de graduação, escolhidos de acordo com as peculiaridades e necessidades da região (nas áreas de gestão e educação), a estrutura do *campus* de Itabaiana recebe diariamente cerca de 2.500 alunos, em três turnos de funcionamento, com concentração no período noturno. Ser universitário é uma conquista familiar da maioria: pesa a responsabilidade de ser o primeiro universitário em uma família de pais que não tiveram oportunidade de acesso à escolarização. Os primeiros exames vestibulares de Itabaiana tiveram uma concorrência superior à concorrência dos mesmos cursos no campus sede da Universidade Federal de Sergipe, trazendo de volta aos estudos alunos que estavam no mercado de trabalho. Ser universitário da Universidade Federal de Sergipe em Itabaiana é um diferencial para esses indivíduos em seus nichos familiares; traz a responsabilidade e o compromisso com o estudo. Durante pelo menos quatro horas por dia, esses estudantes travam contatos próximos, compartilhando valores e conhecimentos (não só dos seus cursos específicos, mas do saber universitário e sua função social). A estrutura da unidade acadêmica – multicursos – facilita o contato e a interação entre todos os alunos; além disso, por muitos virem de cidades circunvizinhas (algumas distantes mais de 50 km de Itabaiana, como é

o caso do município de Carira, na divisa entre Sergipe e Bahia), além do contato na universidade, há o contato durante o trajeto, em ônibus de transporte escolar (muitos subsidiados pelas prefeituras). Esse vínculo entre os universitários fica explícito materialmente por meio da identificação no vestuário: embora não seja obrigatório o uso de uniforme, os graduandos de cada curso se unem e elaboram a camiseta do seu grupo (que pode ser do curso todo, de uma turma do curso ou de parte de uma turma), que é utilizada diariamente como um uniforme, mas com a finalidade de marcar a identidade e o pertencimento ao grupo não entre os pares universitários, mas nas suas redes de relacionamento de origem. O acesso à universidade propicia, também, oportunidades de inserção nos programas institucionais remunerados (iniciação científica, iniciação à docência, extensão, monitoria etc.), de caráter meritocrático. A participação nesses programas é vista como positiva e gera expectativas de continuidade e ascensão nos estudos, alçando os universitários rumo à pós-graduação.

O engajamento social verificado entre os universitários do *campus* de Itabaiana da Universidade Federal de Sergipe nos permite defini-los como constituintes de uma comunidade de prática, nos termos do que propõem Eckert e McConnell-Ginet (2010). Não há um limite geográfico específico para definir essa comunidade, mas um limite de comportamento, preservação e compartilhamento de valores associados ao ser universitário de uma instituição pública no interior. Em termos linguísticos, ser universitário pressupõe a passagem por 11 anos de escolarização formal (ou, em casos de exames supletivos, demonstrar o domínio de conteúdos equivalente a esse tempo de escolarização), em que há um contato direto com a cultura letrada e, por hipótese, o domínio da norma culta da língua. Na sociedade brasileira, o ser universitário está associado ao saber (havia um programa de televisão no formato pergunta-resposta que facultava aos participantes dispor da "ajuda universitária"). Pelo significado social que possui, pelo tipo de exigência do próprio contexto (que prioriza seminários, exposições orais, debates etc. como formas de avaliação), o ambiente universitário é, pois, um dos espaços onde a norma culta falada se manifesta. Ainda que em outros momentos, no seu núcleo familiar, por exemplo, o universitário faça uso de uma variedade linguística marcada por traços estigmatizados (o rotacismo, por exemplo), e ainda que o seu colega também faça uso desse traço, enquanto estiverem desempenhando seus papéis nessa comunidade de práticas universitária, ambos tendem ao monitoramento para a não ocorrência do traço em questão. O banco de dados Falantes Cultos de Itabaiana/SE (CAAE - 0301.0.107.000-11) foi constituído tomando por base o constructo de comunidade de prática. É composto por 20 entrevistas, estratificadas por gênero do entrevistador e do entrevistado e abrangendo a faixa etária dos 19 aos 32 anos, realizadas de acordo com protocolo da entrevista sociolinguística (envolvendo temas como risco de vida, narrativas da infância etc.), mas com o diferencial de que o entrevistador e o entrevistado são membros da comunidade de

prática sob análise, o que faz com que a assimetria seja sensivelmente diminuída. Outro fator importante a ser considerado é que todos os entrevistados e todos os entrevistadores se conhecem, o que não costuma acontecer, por exemplo, em coletas em comunidades de fala.[6] Por haver esse envolvimento, a condução da interação em alguns momentos sai do molde da entrevista sociolinguística, especialmente naqueles em que o tópico discorrido está relacionado com o que constitui a comunidade de prática: o que é ser o primeiro universitário da família, a oportunidade de ser universitário em Itabaiana, prospecções para o futuro. Esses temas são comuns a entrevistado e a entrevistador e, em alguns momentos, a interação assume a forma de desabafo, confidência, compartilhamento, o que é mais uma evidência em favor da constituição efetiva da comunidade de prática.

É importante destacar que o banco de dados não foi constituído para subsidiar, em princípio, estudos ditos de terceira onda. Seu propósito inicial foi dar suporte ao estudo do fenômeno de variação e mudança em categorias verbais do português, em uma perspectiva baseada em frequências de uso.[7] A modelagem da amostra, porém, permite não só estudos de cunho quantitativo, mas também estudos de caráter mais etnográfico, na medida em que as entrevistas são ricas em informações sobre o ser universitário. O banco de dados encontra-se em fase de revisão e em breve será disponibilizado à comunidade científica como mais uma fonte para estudos descritivos de variedades do português falado.

Banco de falares sergipanos

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 proposição do banco de dados *Falares Sergipanos* (CAAE - 0386.0.107.000-11). Por entendermos as ondas propostas por 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 linha de raciocínio, o banco de dados Falares Sergipanos

[6] Deve-se destacar que o corpus do projeto SP2010 é constituído pelo critério "bola de neve" de seleção de informantes (MILROY; GORDON, 2003) – um indica o outro –, procedimento que não se verifica em outros bancos de dados, a exemplo do VARSUL.

[7] Projeto Variação na expressão do tempo verbal passado na fala e escrita de Itabaiana/SE: funções e formas concorrentes (FREITAG, 2009), financiado pela FAPITEC (Edital FAPITEC/FUNTEC-SE Universal 06/2009 Processo n. 019.203.00910/2009-0) e CNPq (Edital MCT/CNPq/MEC/CAPES 02/2010 - Ciências Humanas, Sociais e Sociais Aplicadas Processo n. 401564/2010-0).

(FREITAG, 2011a),[8] em processo de implementação, segue duas linhas de coleta – a de estratificação homogeneizada e a de comunidades de prática.

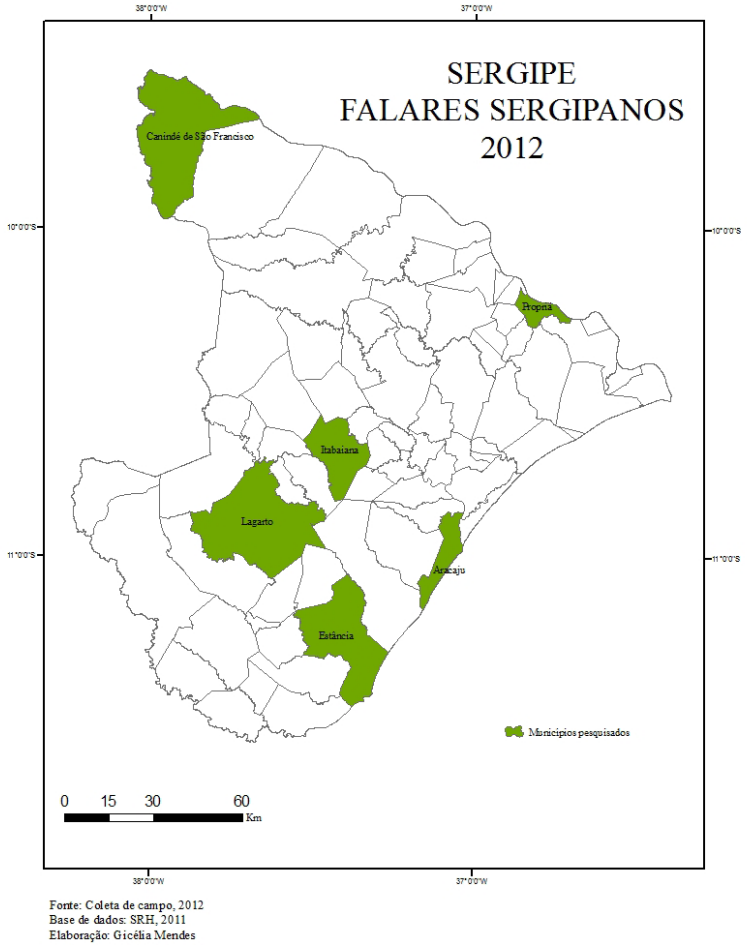
A estratificação homogeneizada é predominante nos bancos de dados já constituídos, como apresentamos em seção anterior. Para o dimensionamento da amostra, foram selecionadas seis cidades representativas do estado de Sergipe, por territórios – Canindé de São Francisco, Itabaiana, Lagarto, Estância, Propriá e Aracaju (figura 1). A estratificação etária dos informantes segue a padronização do IBGE (2010), computando cinco faixas (até 14 anos; 15-24; 25-39; 40-64; mais de 65 anos). A seleção dos informantes (inicialmente dois para cada célula social) seguirá a abordagem "bola de neve", a partir do contato inicial de pesquisador de campo da comunidade, o que será viabilizado pelo fato de a Universidade Federal de Sergipe contar com o curso de Letras na modalidade a distância e a disciplina obrigatória Sociolinguística. Dadas as dimensões do estado e o fato de haver 14 polos universitários, alguns dos quais nas cidades escolhidas para a constituição da amostra, não está prevista, inicialmente, a estratificação por nível de escolarização, o que será feito depois, ao ritmo da coleta, a partir de mapeamento qualitativo. Desse modo, a coleta inicial fornecerá 40 entrevistas por cidade, totalizando 240 entrevistas.

Paralelamente à coleta de estratificação homogeneizada, serão realizadas coletas voltadas para: i) identificação de comunidades de prática, aos moldes do banco de dados Falantes Cultos de Itabaiana/SE, apresentado anteriormente; e ii) observação de efeitos de sexo/gênero, com coletas de dados (entrevistas sociolinguísticas) realizadas por dois entrevistadores: um homem e uma mulher, considerando que Bailey e Tillery (2004) relatam que há estudos que apontam indícios de que o sexo/gênero do entrevistador influencia nos traços linguísticos do entrevistado, propiciando assim análise mais acurada acerca da variação estilística.

Ademais, a partir da identificação de tendências amplas na amostra de estratificação homogeneizada, podem ser realizadas coletas de dados mais particularizadas, que não podem ser definidas aprioristicamente, mas sim a partir da observação empírica de uma realidade.

[8] O projeto Falares Sergipanos integra um projeto maior, intitulado Da expressividade da língua ao mal na literatura: base de pesquisas interinstitucionais do PPGL/UFS, financiado pelo edital CAPES/FAPITEC/SE 06/2012, que, 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), tem como um dos seus objetivos constituir o banco de dados Falares Sergipanos para respaldar a pesquisa da área de concentração Estudos Linguísticos do PPGL/UFS para descrição e estudos comparativos do português e aplicações para o ensino de língua estrangeira e materna.

Figura 1 - Distribuição das cidades constituintes do banco de dados Falares Sergipanos.



Fonte: Freitag (2011a, p.04).

Banco de Dados FALA-Natal

Como já dito, estudos feitos sob a égide da sociolinguística variacionista vêm fomentando a ampliação do conhecimento sobre o português brasileiro desde a década de 1970, através da descrição e da análise de fenômenos variáveis nos âmbitos fonológico, morfológico, sintático, semântico e discursivo. Contudo, há

estados da federação em que tais pesquisas são ainda incipientes ou mesmo inexistentes. É o caso do Rio Grande do Norte, que não conta com um banco de dados de fala com as características necessárias para a pesquisa sociolinguística. Para suprir essa lacuna, Tavares e Martins (2012) propuseram-se a organizar um *corpus* de fala, que será denominado Banco de Dados da Fala do Rio Grande do Norte (FALA-RN) e contará com amostras representativas de diferentes comunidades de fala norte-rio-grandenses. O marco inicial da organização do FALA-RN será a constituição do Banco de Dados FALA-Natal, que congregará entrevistas sociolinguísticas a serem feitas com membros da comunidade de fala do município de Natal, que é a capital e maior centro urbano do estado potiguar. Posteriormente, serão coletadas entrevistas sociolinguísticas em comunidades de fala do interior.

Os informantes do Banco de Dados FALA-Natal serão socialmente estratificados de modo similar a informantes de bancos de dados já existentes no país, a exemplo do PEUL, do VARSUL e do VALPB. Inicialmente, o Banco de Dados FALA-Natal será composto por 48 entrevistas sociolinguísticas com cerca de 60 minutos de duração. Essas entrevistas serão distribuídas, em termos de estratificação social, quanto ao *sexo* (24 informantes de sexo feminino e 24 informantes de sexo masculino), *idade* (12 informantes de 8 a 12 anos, 12 informantes de 15 a 21 anos, 12 informantes de 25 a 50 anos e 12 informantes de mais de 50 anos) e *nível de escolaridade* (12 informantes com ensino fundamental I completo, 12 informantes com ensino fundamental II completo e 12 informantes com ensino médio completo, além de 12 informantes cursando o ensino fundamental I – os indivíduos de 8 a 12 anos). Serão gravados informantes de diferentes bairros das quatro zonas de Natal.

De acordo com o Censo de 2010, a capital norte-riograndense tem, atualmente, 803.739 habitantes (IBGE, 2010). Se considerarmos uma amostragem na condição metodológica ideal, aplicando o corte de 0,5% da população, o banco de dados FALA-Natal deveria contar com 4.019 entrevistas. Em condições reais, o desenvolvimento de um banco de dados com esse número de entrevistas demandaria anos de realização. Com o significativo crescimento da população, se adotássemos tal condição para o desenvolvimento do banco, quando a última entrevista fosse realizada, a comunidade, com certeza, já não seria a mesma. Além disso, a quantidade de informantes também depende de financiamento e de quanto tempo se dispõe para a organização do banco de dados, fatores que, em geral, impedem a coleta de um grande número de entrevistas.

De qualquer forma, um número menor de entrevistas pode ser representativo de tendências gerais da comunidade. Segundo Sankoff (1988, apud TAGLIAMONTE, 2006, p.23), é necessário "[...] não que a amostra seja uma versão em miniatura da população, mas apenas que tenhamos a possibilidade de fazer inferências sobre a população com base na amostra." Cada banco de dados deve ter um

mínimo de representatividade com base em idade, sexo, classe social e/ou nível de educação, o que assegura que a diversidade linguística da comunidade de fala esteja representada na amostra.

Lembramos que a maior coleta de entrevistas sociolinguísticas já feita foi dirigida por Shuy et al., tendo sido gravadas 702 entrevistas em Detroit, nos Estados Unidos. No entanto, as análises mais detalhadas desse *corpus* utilizaram apenas 48 dessas entrevistas, com os informantes distribuídos simetricamente em quatro classes sociais, em um total de 12 informantes por classe (TAGLIAMONTE, 2006).

No caso do Brasil, os bancos de dados costumam ter de dois a três informantes por célula social, o que tende a ser suficiente para a obtenção dos padrões gerais de variação de uma comunidade de fala no que diz respeito a diversos fenômenos variáveis. Quanto ao Banco de Dados FALA-Natal, caso algumas características de uso linguístico variável chamem, por alguma razão, a atenção no conjunto das 48 entrevistas sociolinguísticas iniciais, outras entrevistas poderão ser realizadas – com os mesmos ou outros informantes –, no sentido de possibilitar uma análise mais refinada desses usos.

Uma vez coletadas e armazenadas, as entrevistas integrantes do Banco de Dados FALA-Natal poderão servir de *corpus* para pesquisas que objetivem: i) a descrição e a análise da fala de Natal; ii) a comparação com outros dialetos brasileiros, com o intuito de descrever o português brasileiro de modo mais abrangente e detalhado, e de observar diferenças e semelhanças interdialetais; iii) a comparação com outras vertentes do português, como a europeia; iv) a testagem de teorias linguísticas; v) investigações de natureza social, histórica, antropológica, psicológica, entre outras.

Na constituição do banco de dados FALA-Natal, que está em desenvolvimento,[9] temos nos defrontado com uma série de questões para as quais temos buscado soluções. Entre essas questões, apontamos: i) representatividade da amostra; ii) dificuldade de localização de informantes com certos traços socioeconômicos; iii) necessidade de maior diferenciação de faixas etárias para testar hipóteses relativas à aquisição e à mudança linguística; iv) estratégias para tornar acessíveis à comunidade acadêmica os bancos de dados sociolinguísticos (e as questões éticas aí implicadas); v) validade da comparação de análises realizadas com base em dados extraídos de entrevistas sociolinguísticas feitas recentemente com análises realizadas com base em dados extraídos de entrevistas sociolinguísticas feitas há dez ou vinte anos; vi) como considerar aspectos relacionados à questão da análise estilística pelo viés da terceira onda, nos termos de Eckert.

[9] Estamos em fase de elaboração da estrutura do banco de dados e de seleção de informantes, bem como de treinamento da equipe que realizará as entrevistas. A previsão é que sejam gravadas até março de 2013 as 48 entrevistas que comporão o banco de dados em sua fase inicial.

Em relação a esse último tópico, inicialmente serão coletadas entrevistas sociolinguísticas em uma comunidade de fala ampla – a de Natal – para que seja possível a realização de mapeamentos de tendências gerais de variação e de mudança em relação a essa comunidade, ou seja, em sua primeira fase, o Banco de Dados Fala Natal será composto por entrevistas que permitirão a realização de estudos alinhados à primeira onda da sociolinguística.

Todavia, nossa comunidade de fala alvo abriga, naturalmente, inúmeras comunidades de prática. Com a intenção de aprofundarmos nosso conhecimento acerca das comunidades de prática em que se engajam cada um dos informantes a serem selecionados para o banco de dados, elaboramos uma ficha social a ser preenchida previamente à entrevista na qual constam, entre outras, questões que permitem a obtenção de informações a respeito das diferentes comunidades de prática em que se engaja o informante em sua vida cotidiana. Nessa ficha social, solicitamos, por exemplo, para os informantes de 15 a 21 anos, que respondam às seguintes questões: (i) **Como ocupa seu tempo livre?** e (ii) **Participa de algum grupo (igreja/ jovens/ esporte/ clube)? Se sim, com que frequência?**

Também poderão ser propostos, nas entrevistas, tópicos que estimulem o informante a discorrer sobre as diferentes comunidades de prática das quais faz parte. Com esse fim, elaboramos um roteiro para as entrevistas com sugestões de perguntas que o entrevistador pode fazer ao entrevistado. Entre essas perguntas, estão questões do tipo: **(i) Com quem você passa o tempo, além das pessoas da sua família? O que vocês fazem juntos? Que tipo de lazer vocês têm? (ii) Você participa de algum trabalho voluntário? Como é? (iii) Você participa de algum grupo de jovens? O que vocês fazem juntos? (iv) Você participa de algum grupo da igreja? Como é? (v) Você frequenta algum clube? Qual? Como é? (vi) Algo interessante já aconteceu no clube/grupo de jovens/grupo da igreja quando você estava lá? O que foi? (vii) Descreva o que você faz em um dia, desde que acorda até ir dormir**.

Esse maior conhecimento sobre as comunidades de prática em que se integra cada informante que será obtido com base nas fichas sociais e nas próprias entrevistas poderá ser levado em conta na análise dos fenômenos variáveis. Todavia, as informações presentes nas fichas sociais e nas entrevistas não apenas fornecerão subsídios para uma caracterização mais aprofundada de cada informante no que tange a traços sociais e de prática, como também trarão indícios a respeito de quais comunidades de prática – entre as inúmeras de que participa cada indivíduo – são mais importantes para a realização de estudos nos moldes das segunda e terceira ondas da sociolinguística. Esses indícios fundamentarão as etapas posteriores de construção do Banco de Dados FALA-Natal, em que serão coletadas entrevistas adequadas para contemplar as duas últimas ondas.

Ou seja, para a organização do Banco de Dados FALA-Natal, estamos conscientes da necessidade de organizar não somente um conjunto de entrevistas sociolinguísticas que possibilitem a realização de pesquisas afiliadas à abordagem variacionista alinhada a Labov (um retrato amplo de comunidades de fala definidas geograficamente), mas também à abordagem etnográfica alinhada a Milroy (um retrato local, etnográfico, de comunidades de fala definidas geograficamente) e à abordagem da identidade social alinhada à Eckert (um retrato do(s) indivíduo(s) integrante(s) de comunidades de prática, pelo viés do estilo como elemento central de constituição da *persona*). Pretendemos, pois, num futuro próximo, tornar disponíveis fontes de dados viáveis para pesquisas encaixadas em qualquer uma das três ondas da sociolinguística.

Considerações finais

Como destacamos na introdução, a constituição de um banco de dados sociolinguístico é tarefa dispendiosa e ao mesmo tempo altamente produtiva, por subsidiar estudos de fenômenos variáveis em diferentes níveis linguísticos, com diferentes interfaces teóricas, para vários pesquisadores. Considerando o estado da arte da Sociolinguística no Brasil, e observando a tendência ao direcionamento para estudos de terceira onda (ECKERT, 2012), manifestada já por pesquisadores brasileiros, mas ainda não implementada de forma plena, defendemos que os novos bancos de dados devem, sim, contemplar aspectos relacionados a esta abordagem, mas sem abandonar a tradição consolidada de bancos de dados de estratificação homogeneizada baseados em indicadores sociodemográficos amplos, ditos de primeira onda.

Dado que o interesse em metodologia é uma característica da Sociolinguística e que, como dizem Bailey e Tillery (2004), nem sempre a confiabilidade e a intersubjetividade têm prevalecido nos estudos sociolinguísticos de cunho variacionista, defendemos a necessidade de continuidade de uma metodologia de constituição de *corpus* já consolidada, a fim de permitir a comparação entre amostras, inclusive em tempo real.

FREITAG, R. M.; MARTINS, M. A.; TAVARES, M. A. Brazilian Portuguese sociolinguistic databases and third wave studies: potentialities and limitations. *Alfa*, São Paulo, v.56, n.3 p.907-934, 2012.

- *ABSTRACT*: *Spoken linguistic databases – especially those designed for research with a Variationist Sociolinguistics approach – has been a privileged source for the description of Brazilian Portuguese. In this paper, we discuss methodological procedures that should be adopted in the development of new databases. We trace a short retrospect on already established databases and propose the collection and expansion of corpora from different speech communities – and from different communities of practice. Eckert (2012) describes*

the three analytical practices most commonly embraced by sociolinguistics studies. Each one of these "three waves" (in Eckert's terms) reflects distinct ways of approaching linguistic variation. We suggest strategies to standardize procedures in the development of databases taking into account the three waves of sociolinguistic inquiry, and we shed additional light on the third wave, still incipient in Brazil. The standardization of sociolinguistics databases would make contrastive investigations of different Brazilian dialects easier, contributing, in this way, to the proposition and refinement of generalizations and universal principles of variation and change.

■ *KEYWORDS: Sociolinguistics. Databases. Linguistic variation and change. Social factors. Style.*

REFERÊNCIAS

AMARAL, L. C. *A concordância verbal de segunda pessoa do singular em Pelotas e suas implicações linguísticas e sociais*. 2003. 203f. Tese (Doutorado em Letras) – Universidade Federal do Rio Grande do Sul, 2003.

BAILEY, G.; TILLERY, J. Some sources of divergent data in sociolinguistics. In: FOUGHT, C. *Sociolinguistic variation*: critical reflections. New York: Oxford University, 2004. p.11–30.

BENTES, A. C. Tudo que é sólido desmancha no ar: sobre o problema do popular na linguagem. *Gragoatá*, Niterói, v.27, p.12-47, 2009.

BISOL, L.; MENON, O. P. S.; TASCA, M. VARSUL, um banco de dados. In: VOTRE, S.; RONCARATI, C. (Org.). *Anthony Julius Naro e a linguística no Brasil*: uma homenagem acadêmica. Rio de Janeiro: 7 Letras, 2008. p.50-58.

CAMACHO, R. G. Uma reflexão crítica sobre a teoria sociolinguística. *DELTA*, São Paulo, v.26, n.1, p.141-162, 2010.

CHESHIRE, J.; FOX, S. Was/were variation: a perspective from London. *Language, Variation and Change*, Cambrigde, v.21, p.1-38, 2009.

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

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

ECKERT, P.; MCCONNELL-GINET, S. Comunidades de práticas: lugar onde co-habitam linguagem, gênero e poder (1992). In: OSTERMANN, A. C.; FONTANA, B. (Org.). Linguagem, gênero, sexualidade: clássicos traduzidos. São Paulo: Parábola, 2010. p.93-108.

_____. Communities of practice: where language, gender and power all live. In: COATES, J. (Ed.). *Language and gender*: a reader. Oxford: Blackwell, 1997. p.484-494.

FERRARI, L. V. *Variação linguística e redes sociais no Morro dos Caboclos*. 1994. 204f. Tese (Doutorado em Linguística) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1994.

FREITAG, R. M. K. *Banco de dados de falares sergipanos*. 2011. Projeto de Pesquisa, Universidade Federal de Sergipe, Sergipe, 2011a.

_____. O social da sociolinguística: o controle de fatores sociais. *Diadorim*, Rio de Janeiro, v.8, p.43-58, 2011b.

_____. *Variação na expressão do tempo verbal passado na fala e escrita de Itabaiana/SE*: funções e formas concorrentes. 2009. Projeto de Pesquisa, Universidade Federal de Sergipe, Sergipe, 2009.

FREITAG, R. M. K.; SANTOS, J. C.; SANTOS, S. "Fio do canço": marca linguística identitária do itabaianense. *InterSciencePlace*, [S.l.], v.5, p.1-13, 2009. Disponível em: <http://www.interscienceplace.org/interscienceplace/article/viewArticle/55>. Acesso em: 23 mar. 2012.

GONÇALVES, S. C. L. Projeto ALIP (Amostra Linguística do Interior Paulista). In: MAGALHÃES, J. S.; TRAVAGLIA, L. C. (Org.). *Múltiplas perspectivas em linguística*. Uberlândia: Ed. da UFU, 2008. p.2726-2739. Disponível em: <http://www.filologia.org.br/ileel/artigos/artigo_478.pdf>. Acesso em: 22 mar. 2012.

GUY, G. R. Introdução à análise quantitativa da variação linguística. In: GUY, G. R.; ZILLES, A. M. *Sociolinguística quantitativa*: instrumental de análise. São Paulo: Parábola, 2007. p.19-46.

HORA, D.; WETZELS, L. A variação linguística e as restrições estilísticas. *Revista da ABRALIN*, Brasília, n. esp., p.147-188, 2011. Disponível em: <http://www.abralin.org/revista/RVE1/v4.pdf>. Acesso em: 22 mar. 2012.

IBGE. *Censo demográfico* 2010. Brasília, 2010. Disponível em: <http://www.censo2010.ibge.gov.br>. Acesso em: 22 mar. 2012.

LABOV, W. *Principles of linguistic change*: social factors. Oxford: Blackwell, 2001.

MENDES, R. B. *SP-2010*: construção de uma amostra da fala paulistana. 2011. Projeto de Pesquisa, Universidade de São Paulo, São Paulo, 2011.

MILROY, L.; GORDON, M. *Sociolinguistics*: method and interpretation. Oxford: Blackwell, 2003.

MOLLICA, M. C.; BRAGA, M. L (Org.). *Introdução à sociolinguística*: o tratamento da variação. São Paulo: Contexto, 2004.

MOORE, E. Interaction between social category and social practice: explaining was/were variation. *Language Variation and Change*, Cambrigde, v.22, p.347-371, 2010.

PODESVA, R. J. Phonation type as a stylistic variable: the use of falsetto in constructing a persona. *Journal of Sociolinguistic*, Hoboken, v.11, p.478-504, 2002.

PODESVA, R. J.; ROBERTS, S. J.; CAMPBELL-KIBLER, K. Sharing resources and indexing meanings in the production of gay styles. In: PODESVA, R. J. et al. (Ed.). *Language and sexuality*: contesting meaning in theory and practice. Stanford: CSLI Press, 2002. p.175–90.

SANKOFF, D.; TAGLIAMONTE, S.; SMITH, E. *Goldvarb X*: a variable rule application for macintosh and windows. Department of Linguistics of University of Toronto, Toronto, 2005.

SCHERRE, M. M. P. *Análise e mapeamento de três fenômenos variáveis no português brasileiro*. Projeto de pesquisa, Universidade Federal do Espírito Santo, Vitória, 2011. Disponível em: <http://www.linguistica.ufes.br/sites/www.linguistica.ufes.br/files/Projeto%20de%20Pesquisa%20PPGEL%20-%20UFES%20-%20Marta%20Scherre.pdf>. Acesso em: 22 mar. 2012.

SCHERRE, M. M. P.; RONCARATI, C. Programa de Estudos sobre o Uso da Língua (PEUL): origens e trajetórias. In: VOTRE, S.; RONCARATI, C. (Org.). *Anthony Julius Naro e a linguística no Brasil*: uma homenagem acadêmica. Rio de Janeiro: 7 Letras, 2008. p.37-49.

SEVERO, C. G. A comunidade de fala na sociolinguística laboviana: algumas reflexões. *Voz das Letras*, Concórdia, n.9, p.01-17, 2008. Disponível em: <http://www.nead.uncnet.br/2009/revistas/letras/9/92.pdf>. Acesso em: 22 mar. 2012.

TAGLIAMONTE, S. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press, 2006.

_____. Was/were variation across the generations: view from the city of York. *Language Variation and Change*, Cambrigde, v.10, p.153-191, 1998.

TAVARES, M. A. *Sequenciação de informações na fala de Natal (RN) e de Florianópolis (SC)*: um estudo sociofuncionalista comparativo. 2002. Projeto de Pesquisa, Universidade Federal do Rio Grande do Norte, Natal, 2002.

TAVARES, M. A.; MARTINS, M. A. *Banco de Dados FALA-Natal*: primeiras considerações. Manuscrito. 2012.

WENGER, E. *Communities of practice*: learning, meaning, and identity. Cambridge: Cambridge University Press, 1998.

ZHANG, Q. Rhotacization and the Beijing Smooth Operator: the social meaning of a linguistic variable. *Journal of Sociolingusitic*, Hoboken, v.12, p.201–222, 2008.

Recebido em abril de 2012

Aprovado em julho de 2012