Revista de Estudos da Linguagem, v. 26, n. 2, p. 667-686, 2018

Amostras sociolinguísticas: probabilísticas ou por conveniência?

Sociolinguistic samples: random or convenience?

Raquel Meister Ko. Freitag
Universidade Federal de Sergipe, São Cristóvão, Sergipe / Brasil
rkofreitag@uol.com.br

Resumo: O objetivo deste trabalho é discutir questões relacionadas à amostragem na sociolinguística variacionista, considerando a dimensão probabilística e não probabilística. Conceitos de estatística, como população e amostra, descrição e inferência, são revisados, e os procedimentos de amostragem aleatória e não aleatória são discutidos considerando o viés de seleção e as especificidades da coleta de dados sociolinguísticos para pesquisa de orientação variacionista.

Palavras-chave: sociolinguística; amostragem por cotas; estatística.

Abstract: This paper goal is to discuss sampling in the variationist sociolinguistic approach, both in its random and non-random dimensions. Statistic concepts, such as population and sample, descriptive and inferential statistics are presented; additionally random and non-random sampling procedures are discussed, taking into account the selection bias and the specificity of data collection within variationist sociolinguistics.

Keywords: sociolinguistics; stratified sample; statistics.

Recebido em: 19 de setembro de 2017
Aceito em: 9 de outubro de 2017

eISSN: 2237-2083
DOI: 10.17851/2237-2083.26.2.667-686

1 Introdução

A prática metodológica da sociolinguística variacionista, no Brasil, tem-se pautado tradicionalmente em uma técnica de amostragem dita "aleatória estratificada". Essa técnica consiste em dividir a população por grupos de interesse (células sociais), de modo que todos os falantes[1] pertençam a um e somente um grupo e tenham a mesma chance de ser selecionados. Esse padrão de amostragem, por hipótese, confere confiabilidade e replicabilidade às análises. Neste trabalho, a amostragem é discutida a fim de verificar o quão estratificado é um banco de dados sociolinguísticos.

Inicialmente, são revisados conceitos de estatística, como população e amostra, descrição e inferência. Em seguida, são discutidos os procedimentos de amostragem aleatória e não aleatória, considerando as implicações de escolha (viés) e as especificidades da coleta de dados sociolinguísticos para pesquisa de orientação variacionista.

2 Estatística e tipos de amostra

Dois conceitos básicos em estatística são população e amostra. População refere-se ao conjunto total de elementos; amostra, a um subconjunto dessa população. Com base na amostra, passa-se aos procedimentos de estatística descritiva, que trata da distribuição das frequências um dado fenômeno. A estatística inferencial corresponde ao conjunto de procedimentos que leva à generalização de resultados da amostra para a população. A estatística descritiva responde a perguntas feitas à amostra (quantos? quais?). A estatística inferencial testa hipóteses utilizando informações da amostra para generalizar as respostas para a população.

A sociolinguística variacionista tem evoluído em termos de estatística inferencial: o modelo estatístico para lidar com regras

[1] Na pesquisa sociolinguística, diferentes rótulos têm sido empregados para identificar as pessoas que cedem seu tempo para constituírem amostras linguísticas: informante, falante, sujeito, indivíduo, participante, colaborador, etc. À escolha de um rótulo subjazem matizes do papel que é dado a essa pessoa na pesquisa (embora se refiram a uma pessoa, "informante" e "sujeito" têm cargas semânticas distintas). Fiz a opção por adotar, em todo o texto, o termo "falante", designando a pessoa que fala a língua, ainda que, em alguns contextos, essa escolha resulte em repetições do tipo " o falante falou".

variáveis (CEDEGREN; SANKOFF, 1974; SANKOFF, 1988) vem sendo aprimorado à medida que os níveis de análise vão se ampliando, e novos fatores passam a ser controlados. Hoje, as questões relativas à estatística inferencial da sociolinguística variacionista estão relacionadas à comparação entre as modelagens de efeitos fixos – como as adotadas pelo pacote estatístico VARBRUL e sucessores, como GOLDVARB X (SANKOFF; TAGLIAMONTE; SMITH, 2005) – e efeitos mistos – como RBrul e outros pacotes estatísticos comerciais, como o SPSS – (OLIVEIRA, 2009; JONHSON, 2009, GOMES, 2012; SCHERRE, 2012; GORMAN; JOHNSON, 2013, entre outros), além do uso do pacote estatístico R (R CORE TEAM, 2017) em abordagens da sociolinguística variacionista (TAGLIAMONTE; BAAYEN, 2012; OUSHIRO, 2015, entre outros). O avanço tecnológico, com computadores dotados de processadores mais ágeis e maior capacidade de memória, possibilita a testagem de outros modelos de análise estatística, com a inclusão de mais variáveis e com diferentes níveis de efeitos. O núcleo de discussão não recai sobre o modelo mais apropriado para lidar com a variação linguística (parece ser consenso que modelos de efeitos mistos são mais adequados ao tipo de variáveis que são exploradas na sociolinguística variacionista), mas, sim, sobre o diálogo possível entre as análises nesses últimos 40 anos em modelo de efeitos fixos (pesos relativos). No entanto, é preciso considerar que "muitas das afirmativas estatísticas mais abomináveis são causadas por bons métodos estatísticos aplicados a amostras ruins, e não o contrário" (WHEELAN, 2016, p. 142). Em estatística, costuma-se dizer que, "se entra lixo, sai lixo", não importa o modelo que é adotado na análise; por isso, a atenção aos procedimentos de amostragem é importante.

O foco deste trabalho é justamente tecer reflexões sobre a base para a estatística inferencial da sociolinguística variacionista: a constituição de amostras de fala apropriadas para o estudo da variação linguística. Se, do ponto de vista diacrônico, é preciso fazer o melhor uso de maus dados (LABOV, 1982), do ponto de vista sincrônico, muitas vezes o erro fundamental da análise consiste em uma amostra linguística com vieses: "A análise estatística está em ordem, mas os dados sobre os quais os cálculos são realizados são espúrios ou inadequados" (WHEELAN, 2016, p. 148). Essa é uma questão que vem sendo discutida de modo tangencial em estudos que abordam a metodologia da sociolinguística variacionista.

Com o aforismo do inglês *He can't see the forest for the trees* (em tradução livre, "não se pode ver a floresta pelas árvores"), Guy (2014) discute a correlação entre população e amostra e a possibilidade de generalização. Inferências só são válidas se há à disposição informações mais amplas, não um conjunto seletivo: "Na floresta, quantas árvores você vê e quantas árvores você poderia encontrar? Você encontrou apenas carvalhos ou passou por centenas de bordos para achar seu quinto de carvalhos?" (GUY, 2014, p. 216, tradução minha). Considerando esse aforismo, o plano da estatística descritiva consiste em descrever a floresta, com base em informações das suas árvores. Para extrapolar da parte para o todo, é preciso recorrer à estatística inferencial, mas, para isso, antes, é preciso o corpus da floresta...

2.1 Amostragem aleatória

A estatística inferencial, que generaliza os resultados da amostra para a população, preconiza um processo de amostragem aleatório: a chance de cada um dos falantes ser selecionado para constituir a amostra deve ser a mesma. Exemplo de estudo sociolinguístico com base em amostra aleatória é o Linguistic Atlas of the Gulf States (LAGS) realizado por meio de contatos telefônicos (BAILEY *et al,* 1991).[2] No Brasil, a constituição de uma amostra aleatória simples da comunidade seria possível por meio de uma seleção de falantes, utilizando-se o cadastro de eleitores ou, melhor ainda, recorrendo ao banco de dados constituído por agentes de saúde para fins de cadastro no Programa Saúde da Família, como foi feito no povoado Açuzinho, em Lagarto/SE, que contém a informação de todas as pessoas que efetivamente residem no local e que chega a ser muito mais confiável do que o cadastro de eleitores ou do que as estimativas do IBGE (FREITAG; SANTANA; ANDRADE, 2014). É possível ainda estratificar essa amostra (amostra aleatória estratificada), considerando a proporção de adultos e idosos, homens e mulheres, por exemplo.

A amostragem aleatória por conglomerados explora a existência de grupos em uma dada população. Se esses grupos representam adequadamente a população em relação à característica que queremos medir (os grupos apresentam a variabilidade da população), é possível

[2] No entanto, ainda assim, a aleatoriedade não garante isenção do viés da seleção: a representatividade dos falantes potenciais a serem selecionados resume-se àqueles que possuíam linha telefônica, um bem de consumo de difícil acesso à época da pesquisa.

selecionar um ou mais de um desses conglomerados. Por exemplo, o padrão de comportamento de estudantes de uma escola pública estadual de uma dada comunidade tende a ser estável (estatísticas descritivas providas pela secretaria de educação mostram que os estudantes de escola pública estadual da região metropolitana de Aracaju são oriundos de famílias de uma mesma faixa de renda e estão na mesma faixa etária). Então, estudar o padrão de comportamento de uma dada escola – escolhida aleatoriamente – é como estudar o comportamento de todas as escolas que compõem aquele conglomerado (ver, por exemplo, a amostra Atheneu Sergipense (FREITAG *et al.,* 2016), detalhada na seção 3.2.2).

Raramente é possível realizar amostragens aleatórias.[3] E, para estudos sociolinguísticos de orientação variacionista, essa parece ser a regra.

2.2 Amostragens não aleatórias

As amostras não aleatórias podem ser classificadas em três tipos: por conveniência (acidental), por julgamento (intencional) e por cotas (proporcional), escolhidas por conveniência ou por julgamento.

Em uma amostra por conveniência, o pesquisador de campo seleciona falantes da população em estudo que se mostrem mais acessíveis, colaborativos ou disponíveis para participar do processo, algo do tipo "caiu na rede é peixe". Amostra de julgamento envolve o juízo do pesquisador de campo para selecionar, na população, falantes que sejam boas fontes de informação para os propósitos do processo. A amostragem de cotas prevê um número fixo de falantes em cada uma das categorias, que são preenchidas pelo pesquisador de campo por conta da conveniência e/ou julgamento. Parece ficar bem claro que a técnica de amostragem que predomina na sociolinguística variacionista é esta: a quantidade de falantes das categorias (células sociais) é pré-definida, e o pesquisador de campo vai em busca de falantes disponíveis a participar como voluntários do processo de entrevista sociolinguística (que demanda certo tempo), que sejam representativos da comunidade de fala (que não

[3] É interessante observar como outras áreas das ciências sociais lidam empiricamente com a questão da amostragem. No marketing, Kovacks *et al*. (2004) realizaram análise bibliométrica cujos resultados revelam que 52% dos estudos utilizam amostras de conveniência, 14%, amostras aleatórias simples, 11%, por julgamento, 5%, por cotas e 3%, bola de neve (os outros 15% não especificaram como constituíram suas amostras).

causem estranhamento, que não destoem do padrão da comunidade, etc.). A amostragem não é, portanto, aleatória, pois não são todos os falantes da população que têm igual chance de ser selecionados para a amostra. A amostragem aleatória parte do pressuposto de que o documentador não conhece os sujeitos, o que não ocorre no processo de seleção de falantes para a constituição de amostras sociolinguísticas.[4]

3 Representatividade da amostra

Diferentemente da amostra probabilística (aleatória), a amostra não probabilística apresenta viés amostral. Wheelan (2016) apresenta os vieses amostrais que podem levar a resultados equivocados, dos quais são aplicáveis à abordagem da sociolinguística variacionista o viés de seleção e o viés de publicação.[5]

O viés da seleção incide diretamente na representatividade da amostra. Em propostas para descrever a comunidade em geral, os efeitos do viés de seleção se manifestam, por exemplo, pela influência humana da escolha (sentimentos, afinidades, atitudes, etc.), pela cobertura inadequada da população, pela inabilidade para encontrar certos segmentos da população, pela falta de cooperação em alguns subgrupos (TAGLIAMONTE, 2006).

Considerar a representatividade da amostra barra vieses nos dados que poderiam tornar a generalização impossível (BUCHSTALLER, KHATTAB, 2014). A amostra precisa ser representativa para os propósitos do estudo. Para estudos de cunho sociolinguístico de orientação variacionista, os propósitos costumam estar relacionados à descrição de padrões da comunidade de fala. É nesse ponto que a técnica de amostragem da sociolinguística variacionista começa a

[4] Mesmo quando possível uma amostra aleatória, nem sempre, do ponto de vista da sociolinguística, a aleatoriedade é possível; a distribuição de uma população nunca é geográfica e socialmente aleatória (TAGLIAMONTE, 2006).

[5] Neste texto, trato apenas do viés da seleção. No entanto, o viés de publicação também merece reflexões: achados positivos têm maior probabilidade de ser publicados do que achados negativos. Na abordagem da sociolinguística variacionista, esse viés se manifesta quando os pesquisadores omitem/não informam as variáveis extralinguísticas (preditoras) que foram controladas no modelo, mas que não apresentaram significância estatística. O fato de um fator não ser estatisticamente significativo é sociolinguisticamente significativo!

se distanciar das demais ciências sociais. Uma comunidade de fala é definida não em função de um padrão de uso, mas de um padrão de atitudes (LABOV, 1972); existe um conjunto de atitudes em relação à língua que é compartilhado por quase todos os membros, mas que não necessariamente usam as mesmas formas. Há um viés de intencionalidade para tornar o processo de seleção de falantes compatível com o construto da população: a comunidade de fala.

É também nesse ponto que é importante considerar a distinção entre amostra significativa e amostra representativa: às vezes, um número menor de falantes, quando for possível estabelecer comparação entre grupos, possibilita que se chegue a resultados mais consistentes dos que o recurso de uma única amostra mais numerosa (MARTINS; PINTO, 2015, p. 9). A construção de uma amostra de fala para fins de estudos variacionistas é diferente do modo como é feito nas outras ciências sociais, já que, geralmente, não se pode predizer o quão frequente é uma dada forma/fenômeno linguístico no fluxo da conversação. Uma amostra sociolinguística de orientação variacionista precisa de poucos falantes (20 a 120) cuidadosamente escolhidos para representar a diversidade de comportamentos linguísticos de uma comunidade, com grande volume de material documentado para cada falante (SANKOFF, 2001). A mesma amostra pode ser utilizada para outros estudos, de fenômenos linguísticos diferentes, já que é representativa da estrutura e do uso da fala daquela comunidade.

> A amostra sociolinguística opera na razão inversa das demais ciências sociais: enquanto as ciências sociais operam com amostras com muitos falantes que cedem poucos dados, a sociolinguística opera com amostras com poucos falantes que cedem muitos dados. (SANKOFF, 2001, p. 823)

A conveniência e o julgamento em uma amostra induzem a um viés em relação à população total, produzindo resultado distorcido (LAMEIRÃO, 2014). Manuais de estatística recomendam que sempre que essas técnicas de amostragem são adotadas, os resultados sejam acompanhados por uma descrição detalhada de como a amostra foi obtida, de modo que o leitor possa avaliar qual credibilidade pode dar aos resultados. No caso da amostragem de comunidades de fala, na sociolinguística variacionista, é pertinente incluir os critérios de inclusão e seleção de falantes na metodologia de constituição da amostra,

especificando, por exemplo, de que modo um falante será julgado (por testes de reação subjetiva, por exemplo).

3.1 Amostra por cota fixa ou proporcional

Estratificar uma amostra pressupõe identificar os estratos (células sociais) e calcular a proporção da população de cada estrato representado na amostra. Para o procedimento de estratificação, é preciso considerar as forças sociais que operam sobre a língua (TAGLIAMONTE, 2006), como a classe socioeconômica, o grupo étnico, sexo/gênero, especialmente o papel da mulher (FREITAG, 2015a), idade, com o efeito de pares no grupo (FREITAG, 2005).

Em termos operacionais, a estratificação é implementada em função de características sociodemográficas; algumas podem ser validadas de forma oficial, sem causar constrangimentos, como idade, onde nasceu, o quanto estudou e o sexo (registro civil). Algumas categorizações são mais delicadas, como o quanto ganha (faixa de renda, classe socioeconômica) e o gênero do falante, considerando identificação e orientação. Outras são arbitrárias, feitas pelo pesquisador (à revelia do falante), como ser falante de "português culto" ou "português popular".[6]

A confluência entre os perfis sociais configura as células sociais, ou os estratos, que devem ser preenchidos por falantes que apresentem concomitantemente esses conjuntos de características. Assim, uma amostra hipotética que considere onde mora (centro/subúrbio), sexo civil (masculino/feminino) e idade (jovens e idosos) gera oito estratos, ou células sociais, todas potencialmente ortogonais, ou seja, preenchíveis:[7]

– *Homem, jovem, morador do centro*

– *Homem, jovem, morador do subúrbio*

– *Homem, idoso, morador do centro*

---

[6] Se este critério for especificado previamente, é mais provável que um falante colabore para a constituição de uma amostra de fala culta do que para uma de fala popular, por exemplo.

[7] A quebra da ortogonalidade de uma amostra sociolinguística por cotas se dá, por exemplo, quando se considera a faixa etária e a escolaridade. A célula social para falantes que simultaneamente sejam crianças e universitários tem forte probabilidade de ser vazia (pode existir criança superdotada que curse a graduação antes dos 12 anos, mas é um caso excepcional), quebrando a ortogonalidade da amostra (FREITAG, 2005).

– Homem, idoso, morador do subúrbio

– Mulher, jovem, moradora do centro

– Mulher, jovem, moradora do subúrbio

– Mulher, idosa, moradora do centro

– Mulher, idosa, moradora do subúrbio

Definidas as células sociais, resta a decisão de como preenchê-las: cotas fixas (sempre o mesmo número de falantes em cada célula) ou cotas proporcionais (a proporção de falantes em cada célula corresponde à sua proporção na população). Em bancos de dados brasileiros, o VARSUL, por exemplo, assume uma cota fixa de falantes para cada uma das cidades representadas, independentemente do tamanho da amostra; já o Iboruna assume uma distribuição por cotas proporcionais à população de cada uma das cidades que compõem o banco de dados (FREITAG, 2011; FREITAG; MARTINS; TAVARES, 2012).

A escolha do procedimento de estratificação traz implicações metodológicas. Manter o padrão fixo possibilita comparação com outras amostras, cuja distribuição proporcional pode não ser a mesma. Atribuir proporções (pesos) aos estratos da amostra representa mais fidedignamente a realidade da população.

3.2 Amostras não estratificadas

O relaxamento do rigor da representatividade estatística precisa ser compensado com a convergência de métodos de amostragem de outras disciplinas, como a adoção de modelos de redes sociais e de comunidades de práticas, para garantir diversidade analítica. Para esses modelos, a etnografia é uma etapa necessária, que possibilita captar em que lugar essa língua está (não que isso não seja necessário também nas outras técnicas de amostragem). Na sociolinguística variacionista, o estudo de Penelope Eckert em comunidades escolares de Detroit é o pioneiro a usar esta técnica de abordagem (ECKERT, 1989).

Não há uma metodologia padrão, nem há como fazer um planejamento rigoroso da etnografia de uma dada comunidade; é o acesso do pesquisador de campo e o seu envolvimento na/com a comunidade que vão permitir o desenho da pesquisa. Meyerhoff, Schleef e Mackenzie (2015, p. 59) sugerem que a etnografia da comunidade na

abordagem sociolinguística siga o acrônimo *SPEAKING* proposto por Hymes (2003[1974]): *Settings; Participants; Ends; Act sequence; Keys; Instrumentalities; Norms; Genres*. Em linhas gerais, essa orientação sugere que seja observado como os participantes veem a interação documentada, do ponto de vista físico e psicológico (*settings)*; a descrição dos participantes e e daquilo sobre o que eles falam (*participants*); os objetivos das interações documentadas (*goals*); a forma, o conteúdo e o que acontece nas interações (*act sequence*); o tom, o modo e o estado psicológico da fala (*key*); os registros e as formas da fala (*instrumentalities*), assim como as normas e os gêneros discursivos.

3.2.1 Comunidade de práticas

A comunidade de prática é caracterizada como um agrupamento de falantes (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) (WENGER, 1998; ECKERT; MCCONNELL-GINET, 1997). Estudos de comunidade de práticas não necessitam de amostragem; idealmente, toda a população é considerada, a exemplo do estudo de uma comunidade de práticas religiosas, Praesidium Mãe da Divina Graça da Legião de Maria (católica), situada na zona rural, no povoado Açuzinho, um dos mais de 100 povoados do município de Lagarto, no centro-sul do estado de Sergipe. A documentação sociolinguística dessa comunidade faz parte do banco de dados Falares Sergipanos (FREITAG, 2013) e subsidiou diferentes análises (FREITAG, 2014, 2015b; FREITAG; SANTANA; ANDRADE, 2014, entre outros). O grupo é constituído por 13 participantes, os quais se reúnem sistematicamente duas vezes por semana para tratar das atividades religiosas. As gravações das reuniões e a realização das entrevistas ficaram sob a responsabilidade de Cristiane Conceição Santana e Thais Regina Conceição Andrade. A primeira pesquisadora de campo é residente na localidade, e sua avó foi membro da comunidade de práticas sob análise, o que facilitou o contato e minimizou os efeitos do paradoxo do observador. Paralelamente, foram realizadas entrevistas com vistas a coletar informações acerca da constituição da comunidade, além de investigação documental em atas e livros de registro dessa comunidade; essa investigação possibilitou traçar o perfil da comunidade, fazendo o resgate histórico, e pode ser conferido

em Freitag, Santana e Andrade (2014). Embora no povoado Açuzinho existam mais de 20 grupos religiosos, a escolha não foi aleatória; a acessibilidade foi um dos critérios – talvez o principal – que viabilizou a coleta. Considerando a existência de vários grupos religiosos na mesma comunidade, poderíamos pensar que a amostra de comunidades de práticas é uma amostra por conglomerados. A amostragem por conglomerados explora a existência de grupos em uma dada população. Se esses grupos representam adequadamente a população em relação à característica que queremos medir (os grupos apresentam a variabilidade da população), é possível selecionar um ou mais de um desses conglomerados; foi o que fizemos com a amostra Atheneu Sergipense (FREITAG *et al*., 2016). Essa técnica de amostragem reduz o poder explanatório da análise (não podemos generalizar os resultados obtidos em um grupo de estudantes para a fala de Aracaju), mas garante a replicabilidade (está em andamento a coleta em mais duas escolas, nos mesmos moldes). No entanto, tanto na escolha da amostra da comunidade de práticas religiosa, como na escolha da comunidade de práticas escolares, há um viés de seleção por conveniência – acidentalmente, foi a essas e não a outras que tivemos acesso – e de julgamento – as comunidades de práticas a que tivemos acesso são representativas do padrão de comportamento das demais – configurando uma amostra não aleatória, de composição heterogênea (diferentemente da amostra por cotas).

A composição heterogênea e hierarquizada é uma característica de comunidades de práticas, pois todo agrupamento de pessoas que se reúnem com um propósito comum necessita que alguém sempre esteja à frente para tomar decisões e posicionamentos que favoreçam o progresso da comunidade diante dos objetivos almejados – uma composição mais realista da sociedade o que a estratificação homogeneizada de comunidades de fala.

A comparação dos resultados entre o estudo baseado em amostras de comunidades de fala e de comunidades de práticas torna possível a detecção de padrões de emergência e regularização de variantes na amostra de comunidade 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 (FREITAG; MARTINS; TAVARES, 2012, FREITAG, 2013). O estudo em comunidades de fala possibilita que os resultados sejam aprofundados, desde que se tomem como referência estudos microetnográficos de comunidades de práticas.

3.2.2 Redes sociais/bola de neve

A técnica de amostragem bola de neve, ou amigo do amigo, é um tipo de amostragem utilizado para atingir uma população de difícil acesso ou de baixa incidência de falantes. A rede social, considerando os diferentes hábitos de socialização e o grau de envolvimento com a comunidade local dos falantes iniciais, é utilizada para ter acesso ao coletivo, e cada falante selecionado indica mais um falante (linear) ou dois falantes (exponencial), e assim sucessivamente.

Os laços que ligam cada um dos falantes podem ser de primeira ordem (falantes que diariamente estão interagindo), ou de segunda ordem, (falantes que se interligam indiretamente). Redes são caracterizadas também quanto à sua densidade e "plexidade". Quando todos os membros se conhecem, a rede é de alta densidade; quando não há o contato entre todos os membros, a rede é de baixa densidade. Em relação à plexidade, os membros podem estabelecer laços multiplex, ou seja, duas pessoas se relacionam em mais de um papel social e estão presentes em mais de um grupo, e laço uniplex, quando o laço entre duas pessoas é baseado em apenas um relacionamento.

Na sociolinguística variacionista, o modelo de rede social foi adotado no estudo de Milroy (1980), em três comunidades de classe trabalhadora (duas católicas e uma protestante) em Belfast, Irlanda, que examinou diferentes tipos de redes, dentro das quais os falantes se socializavam, e a correlação da força da rede com variáveis linguísticas. Para medir a força da rede, Milroy (1980) propôs uma combinação de traços para controlar multiplexidade e densidade da rede, baseada em uma escala de seis pontos, do 0 a 5, controlando os seguintes parâmetros: se o falante faz parte de uma rede territorialmente constituída (rede densa), se tem laços fortes de parentesco (rede multiplexa), se trabalha no mesmo lugar com ao menos dois outros membros da mesma comunidade (rede multiplexa), se compartilha o mesmo local de trabalho com ao menos dois outros membros do mesmo sexo da mesma área (rede multiplexa), se desenvolve trabalhos voluntários nas horas vagas (rede multiplexa).

Na perspectiva da sociolinguística brasileira, a abordagem de redes tem sido adaptada e utilizada na seleção de falantes para as amostras

estratificadas considerando a plexidade e a densidade (BATTISTI, 2014; ARAUJO; SANTOS; FREITAG, 2014).[8]

4 Tamanho da amostra

O tamanho da amostra, para fins da generalização da estatística inferencial, requer que sejam consideradas a margem de erro (probabilidade de o intervalo conter a média verdadeira) e a significância (grau de acurácia para que determinado resultado seja considerado válido) em amostras aleatórias. Em amostras não aleatórias, o tamanho envolve a decisão de quantos falantes vão preencher cada cota/célula social. Normalmente, essa decisão envolve disponibilidade de tempo e recursos do pesquisador. No exemplo da seção 3.1, se as oito células sociais forem preenchidas por cotas de um falante, a coleta necessitará de oito falantes. Trabalhar com o número mínimo não é uma situação adequada, pois, ainda que haja um teste de julgamento para incluir ou não o falante na amostra, sem um parâmetro da célula. Fica difícil julgar. Então, vamos aumentar a cota para dois, o que leva a uma amostra de 16 falantes. Dois falantes é um número mínimo para a constituição de amostras sociolinguísticas por cotas fixas; no entanto, podemos ampliar a amostra para garantir representatividade. Com cinco falantes por cota, a amostra necessitará de 40 falantes. E assim sucessivamente.

No entanto, a amostra sociolinguística opera na razão inversa das demais ciências sociais (SANKOFF, 2001), operando com amostras com poucos falantes que cedem muitos dados. A depender dos recursos e disponibilidade, pode ser realizada ampla coleta, mas tratamento estatístico de apenas uma parte dos dados. E é nesse ponto que é preciso considerar o fenômeno linguístico sob análise. Muito mais importante do que a cota por célula é garantir a representatividade do fenômeno

[8] Com base na proposta de Blake e Josey (2003), Oushiro (2011) e Araujo, Santos e Freitag (2014) desdobram-se critérios para controle de densidade e plexidade da rede de falantes: *Grau* 1 – Bastante próximo. Os falantes têm laços fortes (amizade, parentesco, colega de trabalho ou escola etc.) e interagem diariamente; *Grau* 2 – Próximo. Os falantes interagem frequentemente, mas não têm laços fortes; *Grau* 3 – Próximo. Os falantes não interagem frequentemente e não têm laços fortes; *Grau* 4 – Neutro. Os falantes se conhecem, mas não interagem com frequência; *Grau* 5 – Distante. Os interlocutores não se conheciam anteriormente e só conversaram no momento da gravação da interação.

linguístico no modelo de análise construído. Meyerhoff, Schleef e Mackenzie (2015) recomendam que, para garantir resultados confiáveis e acurados quando submetidos ao tratamento estatístico, cada fator tenha, no mínimo 30 ocorrências por células. Isso implica dizer que, após a estratificação social da amostra (8 células sociais), é preciso computar as células das outras variáveis preditoras dependentes e garantir que o tamanho da amostra possibilite identificar pelo menos 30 ocorrências para cada variável preditora independente. Em fenômenos fonológicos, isso é possível com o número mínimo de falantes por cota. Já em fenômenos sintáticos mais raros, esse dimensionamento requer ou mais horas de fala por falante, ou mais falantes por cotas. Em termos de nível de rigor e critério, não existe estudo perfeito, não existem condições ideais. Cada fenômeno e cada realidade impõem restrições e dimensionamentos específicos. Por conta disso, amostras não probabilísticas não possibilitam avaliar a precisão do resultado.

O quantitativo de ocorrências necessário para uma análise de um fenômeno levanta a questão dos custos: o desenho, a coleta e o armazenamento de uma amostra linguística envolvem recursos humanos altamente especializados (treinados não só para a pesquisa de campo e abordagem de falantes, mas também para os procedimentos de transcrição do áudio e anotação dos dados), o que implica recursos financeiros. Os órgãos de financiamento da pesquisa sociolinguística no Brasil – assim como ocorre nas demais áreas da ciência – têm valorizado projetos que possibilitem o compartilhamento de amostras por mais bancos de dados sociolinguísticos, que possam ser utilizados mais de uma vez e por mais pesquisadores, para estudar diferentes fenômenos (FREITAG, 2016, 2017).

5 Comparabilidade versus realibilidade

Considerando os aspectos de amostragem discutidos, o leitor que chegou a este ponto do texto pode se perguntar se o que vem sendo feito não tem validade. A resposta é, definitivamente, sim! A tarefa de constituição de bancos de dados é dispendiosa, mas, acima de tudo, é irreplicável temporalmente. Uma vez feita a coleta, não é possível voltar no tempo para corrigir os erros de amostragem que porventura tenham ocorrido. Daí a importância de um planejamento, considerando o objeto do estudo (sua recorrência) e os recursos disponíveis (pessoas

que estarão envolvidas na coleta dos dados, tempo disponível para os procedimentos de coleta e armazenamento dos dados, infraestrutura e equipamentos disponíveis).

Na constituição de novos bancos e na expansão dos bancos já existentes, é desejável seguir o padrão de estratificação já convencionalizado e difundido, o que possibilita a comparação de resultados. É possível revisar o dimensionamento amostral, a fim de garantir a reabilidade, ou seja, a consistência da aplicação de métodos estatísticos após a sua repetição. No entanto, a comparabilidade das amostras, garantindo a série histórica, tem primazia em relação à reabilidade estatística (FREITAG; ROST-SNICHELOTTO, 2015).

6 Conclusão

O tipo de amostragem que tem sido utilizado em estudos sociolinguísticos de orientação variacionista, de fato, não é probabilística aleatória estratificada, e, sim, de cotas por conveniência e julgamento, na medida em que os falantes são selecionados pelo critério de disponibilidade e voluntariedade em aceitar os termos da coleta, especialmente as amostras que são chanceladas por Comitê de Ética em Pesquisa (FREITAG, 2017). A conveniência possibilita a operacionalidade da coleta, mas impõe à análise menor poder explanatório; por não atender a um critério estatístico, não pode (ou, melhor, não deve) ser generalizada a uma população. Amostras assim constituídas não poderiam, em tese, subsidiar generalizações sobre "a" língua falada em tal lugar por não garantirem a representatividade da população. E, por serem pautadas na conveniência, limitam a replicabilidade, na medida em que há um viés de seleção.

Essa opção metodológica levanta questões relacionadas à generalização dos resultados e o poder explanatório da estatística inferencial subjacente ao modelo de análise utilizado: o quão acurada é a representação da população na amostra? o quão generalizáveis são os resultados? Nossa prática se pauta pelo dimensionamento de tempo e recursos e não necessariamente pela representatividade da amostra. Muitas vezes temos que fazer bom uso de "maus dados".

Assim, cabe a recomendação de manuais de estatística: em estudos com amostragem por conveniência, os resultados devem acompanhar uma descrição detalhada da metodologia de obtenção da amostra para permitir ao leitor o juízo de credibilidade da análise.

Agradecimentos

Este texto foi debatido no encontro do GT de Sociolinguística da Anpoll, ocorrido durante o 31º Enanpoll, em 2016. Agradeço aos debatedores pelos comentários e pela audiência, especialmente a Livia Oushiro, Rosane Andrade Berlinck, Marco Antonio Martins e Silvia Rodrigues Vieira, assim como aos pareceristas anônimos da Relin, que contribuíram significativamente para o aprimoramento do texto.

Referências

ARAUJO, Andréia Silva; SANTOS, Kelly Carine; FREITAG, Raquel Meister Ko. Redes sociais, variação linguística e polidez: procedimentos de coleta de dados. In: FREITAG, Raquel Meister Ko. (Org.). *Metodologia de coleta e manipulação de dados em Sociolinguística*. São Paulo: Blücher, 2014. p. 99-116.

BAILEY, Guy; WIKLE, Tom; TILLERY, Jan; SAND, Lori. The apparent time construct. *Language, Variation and Change*, Cambridge, v. 3, n. 3, p. 241-264, 1991.

BATTISTI, Elisa. Redes sociais, identidade e variação linguística. In: FREITAG, Raquel Meister Ko. (Org.). *Metodologia de coleta e manipulação de dados em Sociolinguística*. São Paulo: Blücher, 2014. p. 79-98.

BLAKE, Renée; JOSEY, Meredith. The/ay/diphthong in a Martha's Vineyard community: What can we say 40 years after Labov? *Language in Society*, Cambridge, v. 32, n. 4, p. 451-485, 2003. DOI: 10.10170S0047404503324017

BUCHSTALLER, Isabelle; KHATTAB, Ghada. Population samples. In: PODESVA, Robert; SHARMA, Devyani (Ed.). *Research methods in linguistics*. Cambridge: Cambridge University Press, 2014, p. 74-95.

CEDERGREN, Henrietta; SANKOFF, David. Variable rules: Performance as a statistical reflection of competence. *Language*, Washington, v. 50, n. 2, p. 333-355, 1974. DOI: 10.2307/412441

ECKERT, Penelope. *Jocks and burnouts*: Social categories and identity in the high school. Michigan: Teachers College Press, 1989.

ECKERT, Penelope; MCCONNELL-GINET, Sally. Communities of practice: where language, gender and power all live. In: COATES, Jennifer. (Ed.). *Language and Gender:* a reader. Oxford: Blackwell, 1997. p. 484-494.

FREITAG, Raquel Meister Ko. (Re)discutindo sexo/gênero na sociolinguística. In: FREITAG, Raquel Meister Ko.; SEVERO, Cristine Gorski. (Org.). *Mulheres, linguagem e poder*: estudos de gênero na sociolinguística brasileira. São Paulo: Editora Edgard Blücher, 2015a. p. 17-74.

FREITAG, Raquel Meister Ko. Banco de dados falares sergipanos. *Working Papers em Linguística*, Florianópolis, v. 14, n. 2, p. 156-164, 2013. DOI: http://dx.doi.org/10.5007/1984-8420.2013v14n2p156

FREITAG, Raquel Meister Ko. Covariação em uma comunidade de práticas. In: LOPES, Norma da Silva; RAMOS, Jânia; OLIVEIRA, Josane Moreira. (Org.). *Diferentes olhares sobre o português brasileiro*. Feira de Santana: Editora UEFS, 2014. p. 13-30.

FREITAG, Raquel Meister Ko. Desafios teóricos-metodológicos da sociolinguística variacionista. In: PARREIRA, Maria Cristina *et al*. (Org.). *Pesquisas em Linguística no século XXI*: perspectivas e desafios teórico-metodológicos. São Paulo: Cultura Acadêmica, 2015c. v. 27, p. 29-43.

FREITAG, Raquel Meister Ko; SNICHELOTTO, Cláudia Andrea Rost. Análises contrastivas: estabilidade, variedade ou metodologia?. *Working Papers em Linguística*, Florianópolis, v. 16, n. 1, p. 157-169, 2015. DOI: http://dx.doi.org/10.5007/1984-8420.2015v16n1p157

FREITAG, Raquel Meister Ko. *et al*. Avaliação e variação linguística: estereótipos, marcadores e indicadores em uma comunidade escolar. In: FREITAG, Raquel Meister Ko.; SEVERO, Cristine Gorski; GÖRSKI, Edair Maria. *Sociolinguística e política linguística*: olhares contemporâneos. São Paulo: Blucher, 2016. p. 141-160.

FREITAG, Raquel Meister Ko. Idade: uma variável sociolinguística complexa. *Línguas & Letras*, Cascavel, v. 6, n. 11, p. 105-121, 2005. DOI: http://dx.doi.org/10.5935/rl&l.v6i11.875

FREITAG, Raquel Meister Ko. O social da sociolinguística: o controle de fatores sociais. *Diadorim*, Rio de Janeiro, v. 8, p. 43-58, 2011.

FREITAG, Raquel Meister Ko. Socio-stylistic aspects of linguistic variation: schooling and monitoring effects. *Acta Scientiarum. Language and Culture*, Maringá, v. 37, p. 127-136, 2015b. DOI: http://dx.doi.org/10.4025/actascilangcult.v37i2.24240

FREITAG, Raquel Meister Ko.; MARTINS, Marco Antonio; TAVARES, Maria Alice. Bancos de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: potencialidades e limitações. *Alfa*, Araraquara, v. 56, n. 3, p. 917-944, 2012. DOI: http://dx.doi.org/10.1590/S1981-57942012000300009

FREITAG, Raquel Meister Ko. Sociolinguística no/do Brasil. *Cadernos de Estudos Linguísticos*, Campinas, v. 58, n. 3, p. 445-460, 2016.

FREITAG, Raquel Meister Ko. *Documentação Sociolinguística, coleta de dados e ética em pesquisa*. São Cristóvão: EdUFS, 2017.

FREITAG, Raquel Meister Ko.; SANTANA, Cristiane Conceição; ANDRADE, Thaís Regina Conceição. Práticas constitutivas do Povoado Açuzinho. *Ambivalências*, São Cristóvão-SE, v. 2, p. 194-217, 2014. DOI: http://dx.doi.org/10.21665/2318-3888.v2n3p194-217

GOMES, Christina Abreu. Para além dos pacotes estatísticos Varbrul/Goldvarb e Rbrul: qual A concepção de gramática? *Revista do GELNE*, Natal-RN, v. 14, n. 1/2, p. 257-272, 2012.

GORMAN, Kyle; JOHNSON, Daniel Ezra. Quantitative analysis. In: BAYLEY, Robert; CAMERON, Richard; LUCAS, Ceil (Ed.). *The Oxford handbook of sociolinguistics*. Oxford: Oxford University Press, 2013. p. 214-240. DOI: 10.1093/oxfordhb/9780199744084.001.0001

GUY, Gregory. Words and numbers: statistical analisys in sociolinguistics. In: HOLMES, Janet; HAZEN, Kirsk (Ed.). *Research methods in sociolinguistics*: a practical guide. Malden: Wiley & Sons, 2014. p. 194-210.

HYMES, Dell. *Foundations in sociolinguistics*: An ethnographic approach. London: Routlege Press, 2003.

JOHNSON, Daniel Ezra. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, Wiley Online Library, v. 3, n. 1, p. 359-383, 2009. DOI: 10.1111/j.1749-818X.2008.00108.x

KOVACS, Michelle Helena *et al*. Podemos confiar nos resultados de nossas pesquisas? Uma avaliação dos procedimentos metodológicos nos artigos de marketing do EnANPAD. In: EMA - ENCONTRO DE MARKETING DA ANPAD, I, 2004. Porto Alegre. Anais… Porto Alegre: Anpad, 2004. p. 1-15.

LABOV, William. Building on empirical foundations. In: LEHMANN, Winfred; MALKIEL, Yakov (Ed.). *Perspectives on historical linguistics*. New York: John Benjamins Publishing, 1982. p. 17-92.

LABOV, William. *Sociolinguist patterns*. Philadelphia: University of Pennsylvania Press, 1972.

LAMEIRÃO, Adriana Paz. O controle metodológico como meio para assegurar a credibilidade de uma pesquisa de survey. *Pensamento Plural,* Pelotas, n. 14, p. 41-63, 2014.

MARTINS, Fernanda; PINTO, Maria da Graça Lisboa Castro. Procedimentos de pesquisa: alguns conselhos práticos para o estudo também psicolinguístico de realidades concretas. *Letras de Hoje*, Porto Alegre, v. 50, n. 1, p. 7-12, 2015. DOI: https://doi.org/10.15448/1984-7726.2015.1.20569

MEYERHOFF, Miriam; SCHLEEF, Erik; MACKENZIE, Laurel. *Doing Sociolinguistics*: A practical guide to data collection and analysis. New York: Routledge, 2015.

MILROY, Lesley. *Language and social networks*. Oxford: Blackwell, 1980.

OLIVEIRA, Alan Jardel. Análise quantitativa no estudo da variação linguística: noções de estatística e análise comparativa entre Varbrul e SPSS. *Revista de Estudos da Linguagem*, Belo Horizonte, v. 17, n. 2, p. 93-119, 2009. DOI: http://dx.doi.org/10.17851/2237-2083.17.2.93-119

OUSHIRO, Livia. *Uma análise variacionista para as interrogativas-Q*. 2011. 160 f. Dissertação (Mestrado em Linguística) – Universidade de São Paulo, 2011.

OUSHIRO, Lívia. *Identidade na pluralidade:* avaliação, produção e percepção linguística na cidade de São Paulo. 2015. 394 f. Tese (Doutorado em Linguística) – Universidade de São Paulo, 2015.

R CORE TEAM (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL:< https://www.R-project.org/.>

SANKOFF, David. Statistics in sociolinguistics. In: MESTHRIE, Rajend (Ed.). *Concise Encyclopedia of Sociolinguistics*. New York: Elsevier, 2001. p. 828-834.

SANKOFF, David. Variable rules. In: AMMON, Ulrich; DITTMAR, Norbert; MATTHEIER, Klaus J. (Ed.). *Sociolinguistics*: An international handbook of the science of language and society, Berlin: Walter de Gruyter, 1988. v. 2, .

SANKOFF, David; TAGLIAMONTE, Sali; SMITH, Eric. *GoldVarb X*: Variable Rule Application for Macintosh and Windows. Toronto: University of Toronto, 2005.

SCHERRE, Maria Marta Pereira. Padrões sociolinguísticos do português brasileiro: a importância da pesquisa variacionista. *Tabuleiro de Letras*, Salvador, n. 4, p. 1-32, 2012.

TAGLIAMONTE, Sali. *Analyzing Sociolinguistic Variation*. Cambridge: Cambridge University Press, 2006.

TAGLIAMONTE, Sali A.; BAAYEN, R. Harald. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language, Variation and Change*, Cambridge, v. 24, n. 2, p. 135-178, 2012. DOI: https://doi.org/10.1017/S0954394512000129

WENGER, Etienne. *Communities of practice*: learning, meaning, and identity. Cambridge: Cambridge University Press, 1998. DOI: https://doi.org/10.1017/CBO9780511803932

WHEELAN, Charles. *Estatística:* o que é, para que serve, como funciona. Rio de Janeiro: Zahar, 2016.