A dadidade (ou dadidão) do dado

Raquel Meister Ko. Freitag

FREITAG, Raquel M.K. A dadidade (ou dadidão) do dado, *Linguística Rio*, vol.3, n.1, maio de 2017.

ISSN: 2358-6826

Informações do autor
Raquel Meister K.o Freitag
Depto. Letras Vernáculas e PPG Letras, Universidade Federal de Sergipe
Contato: rkofreitag@uol.com.br

Outras informações
Enviado: 19 de fevereiro de 2017
Aceito: 18 de abril de 2017
Online: 02 de junho de 2017

RESUMO: O objetivo deste texto é tecer considerações sobre a entidade "dado linguístico", a fim de contribuir para a formação do linguista. São abordadas questões relativas à organização de bancos de dados linguísticos, considerando autoria, posse e propriedade dos dados e a aleatoriedade dos dados.

PALAVRAS CHAVE: Dados linguísticos, Autoria, Aleatoriedade.

Introdução

São poucos os cursos de graduação em Linguística no Brasil (conheço três, todos no estado de São Paulo).[1] Enquanto a formação de engenheiros, arquitetos e geógrafos é feita em um curso de graduação em específico, no Brasil, a maioria dos linguistas advém de cursos de graduação em Letras, na habilitação licenciatura, mas tem sua formação verticalidade na área em pequenos nichos de grupos de pesquisa (quando vinculados a um projeto de pesquisa na iniciação científica ou programa equivalente), ou, então, fazem-na por conta própria. A formação na academia, via inserção em grupos de pesquisa, tende a levar os jovens pesquisadores a se espelharem nos valores e no modo de fazer pesquisa dos sêniores, como que um efeito do *habitus* (BOURDIEU, 1974), o que leva a situações em que, muitas vezes, fazemos as coisas de um jeito porque sempre se fez assim, nem

[1] Na pós-graduação, há bem mais cursos de Linguística, embora predominem os cursos de Letras, com áreas de concentração/linhas específicas para estudos linguísticos, especialmente nas regiões Norte e Nordeste. No entanto, mesmo nos programas de Linguística, a adequação aos ditames da Capes quanto à organização curricular e prazos deixa pouco espaço para aspectos mais práticos da formação. Na minha experiência, tanto como aluna de graduação e pós-graduação, como orientadora de graduação e de pós-graduação, observo que a formação metodológica que prevalece na trajetória do pesquisador é aquela advinda do contato com a graduação.

sempre se sabe por quê, só se sabe que é assim. Ao contrário do que se esperaria, na academia, em geral, há muito pouco espaço para a reflexão metodológica e, por conseguinte, para o rompimento das práticas já consagradas, gerando inovação[2].

Por outro lado, o linguista autodidata esbarra em um problema: no Brasil, temos muito nos dedicado à apresentação de teorias, interfaces e releituras de teorias, mas não temos muita discussão sobre a metodologia que dá suporte a essas teorias todas. No máximo, encontramos descrições de métodos de pesquisa sucintas, na voz passiva, dando a falsa impressão de que é extremamente fácil executar um procedimento metodológico nas diferentes áreas da Linguística.

Uma chamada de publicação que tem como objetivo trazer à discussão a formação do linguista é uma oportunidade para levantar pontos que ainda são falhos. E um dos pontos que considero crucial e pouco refletido de forma consciente é a dadidão ou a dadidade do dado.

Dadidão e dadidade não são palavras, insiste em me lembrar o corretor do editor de texto. Para explicar o que eu quero dizer com esses neologismos, primeiro preciso discutir a sua raiz, o verbo *dar*.

Qual é a melhor análise para o verbo *dar* no português brasileiro? É um verbo ditransitivo? Verbo leve ou suporte? São diversas as possibilidades para fomentar uma investigação linguística – ver por exemplo os estudos de Silva (2005) e Scher (2006). O verbo *dar* assume conotações mais vulgares, quando se torna intransitivo (ou, ao menos, com os objetos elípticos), ou pode ainda ser hipercorrigido, como é o caso de uma propaganda de xampu, que dá brilho aos cabelos (fig. 1)

[2] Não é por comodismo, não é por aculturação. Eu ousaria dizer que fazemos o que sempre fazemos porque sabemos a que resultados vamos chegar. E precisamos chegar a um resultado para poder dar propulsão à mola do financiamento: tenho que ter X artigos publicados em revistas para poder ter chances de concorrer a um edital de fomento, ou simplesmente para ter progressões na carreira. Corre-se menos riscos ao se tentar publicar um artigo com um tema abordado ao modo do que que se espera no campo do conhecimento do que um artigo resultante do teste de novas abordagens (que podem ou não ter resultados publicáveis).
Se estamos tratando da **formação** do linguista, é preciso trazer à tona questões que são delicadas, mas que precisam ser discutidas, como a naturalização de comportamentos na academia (apenas para situar: https://www.cartacapital.com.br/revista/929/a-academia-e-seus-comportamentos-patologicos).

Figura 1: Foto do xampu doador de brilho.

Qual seria a forma deverbal do verbo dar? Emília, do Sítio do Pica-Pau Amarelo de Monteiro Lobato (2014 [1933]), era uma "dadeira de ideias"; por que não um xampu "dador de brilho"? *Dar* e *doar* apresentam acepções muito próximas, mas não são a mesma coisa. Eu me arriscaria a dizer que não têm o mesmo valor de verdade, por isso não podem ser considerados como variantes de uma mesma variável linguística. Porque em Sociolinguística, analisamos dados, e não doados! E é sobre essa entidade, o dado linguístico, que vou tratar neste texto.[3] Inicialmente, discuto o que é a entidade dado linguístico, para, em seguida, encaminhar pontuações sobre a dadidade/dadidão dos dados na Sociolinguística, mas que são extensíveis a outras abordagens, não só da Linguística.

1. O que é um dado?

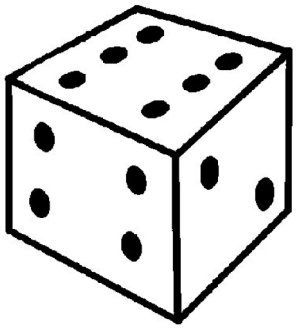

Figura 2: Um dado.

Na nossa cultura, concretamente, dado é uma forma geométrica tridimensional com seis lados, utilizada em jogos (fig. 2). Mais uma faceta do *dado*: sua relação com a sorte, com o acaso (e com a aleatoriedade dos dados). E *dado* assume outra significação: um qualquer. Isso permite construções como "dado dado". *Dado* é um termo técnico polissêmico em Linguística, e uma expressão como "Um dado dado" tem mais de uma interpretação.

[3] Possenti (2009[1998]) utiliza o recurso gráfico do negrito e do itálico para representar a diferença entre duas acepções de dado na Análise do Discurso (o dado *dado* e o dado **dado**). Um sentido seria o dado que é herdado, doado, mais ou menos como o xampu, e o outro sentido seria o dado que realmente é dado, que resolve, dado bom mesmo. Confesso que não consigo entender essas filigranas nesse campo, mas registro a fonte da distinção entre os dados, encaminhando para meu campo, que é de orientação mais funcional.

1. Um dado **dado.**
2. Um **dado** dado**.**

O negrito nos excertos marca a posição de núcleo do sintagma nominal. Em (1), o *dado* é um qualquer, no sentido de escolhido ao acaso, e, em (2) o *dado* é dado (no sentido de doado, como no xampu). A posição de núcleo ou adjunto do sintagma ocupada por *dado* pode, por exemplo, distinguir as abordagens empíricas ou intuitivas, em um viés aos moldes da discussão entre formalismo vs. funcionalismo no debate travado na revista Delta nos anos 1990.[4] A oposição "dado **dado**" e "**dado** dado", neste debate, distinguiria o linguista de gabinete, que cria, pensa um dado para ilustrar seu fenômeno (mas que poderia ser qualquer outro dado), ao linguista de campo, que vai em busca de um dado que saia de algum lugar, que seja dado/doado. Uma distinção entre esses dois tipos de dado poderia ser sistematizada na oposição dado induzido vs. dado espontâneo. Mas, como espero explicitar convincentemente mais à frente, esta é uma distinção falaciosa e falseável. Para conduzir a discussão que proponho, vou assumir a distinção de dado na posição de núcleo do sintagma; como em uma abordagem da sociolinguística de orientação variacionista, vou discutir se as formas de base dos deverbais de *dar* e *doar* são intercambiáveis em outro contexto: na lexia *banco de dados*.

2. O que é um banco de dados?

As abordagens empíricas que visam descrever os usos linguísticos a partir das regularidades prescindem de um suporte: os dados linguísticos. Os dados estão aí, os dados estão no mundo. Podemos dizer, sem medo de errar, que os dados são gerados espontaneamente. Mas ainda não existe uma máquina que pegue dados no mundo, organize-os e os coloque à disposição de um pesquisador, o que revela um grande problema da nossa área: a pouca interação com áreas tecnológicas, que lidam também com processamento

[4] Considero essencial para a formação de um linguista, seja qual for a sua orientação teórica, conhecer o debate na revista Delta: Votre e Naro (1989), Nascimento (1990), Dillinger (1991) e Kato (1998). Houve um tempo em que publicações eram lidas e debatidas, replicadas, treplicadas e muitos aprendiam com a exposição de diferentes pontos de vista sobre fatos da língua. Não sei dizer em que momento ou por que perdemos esse hábito de ler, discutir e compartilhar nossas impressões sobre os trabalhos em desenvolvimento. Só sei dizer que perdemos muito com essa não prática.

da linguagem, como a linguística computacional que é desenvolvida nos cursos de Ciências da Computação. Afinal, se Siri, da Apple, Cortana, da Microsoft, e Allo, da Google, assistentes pessoais, "entendem" seus proprietários, inclusive em português e com certa precisão, é sinal de que existe tecnologia para captar estímulos em áudio e os transformar em registro escrito, a serem processados por algoritmos de *parsing* (a transcrição automática, um sonho de todos que trabalham com dados de áudio!). No entanto, ainda estamos distantes de nos valer destas tecnologias para nossas pesquisas, pelo menos no que se tem feito em Sociolinguística, e no Brasil (e este é outro ponto que linguistas em formação deveriam investir: em formas de automatizar o trabalho, para podermos dar conta de um volume muito mais amplo de dados, e nos dedicar mais tempo à análise).

Mas, mesmo assim, por mais tecnologia que exista (e que potencialmente estivéssemos utilizando), ainda resta todo um trabalho braçal que só pode ser feito por um humano, especialmente na coleta de dados clássica da Sociolinguística Variacionista, com entrevistas gravadas com falantes. A coleção de dados coletados (*corpus* ou amostra) é organizada e constitui um banco de dados. Para contribuir com a formação do pesquisador, gostaria de trazer para a discussão aspectos relacionados aos bancos de dados sociolinguísticos relativos ao trabalho subjacente à organização dos dados e as definições de autoria, posse e propriedade dos produtos derivados deste processo e às questões de aleatoriedade dos dados.

Para isso, vou pedir licença para: a) sair do padrão acadêmico-científico, com menos rigor científico e mais teor anedótico, a fim de me aproximar mais com o público; e, em decorrência disso, b) me eximir da responsabilidade de contar o nome dos santos dos milagres relatados (embora em alguns casos isso seja impossível de acontecer, dada a repercussão geral).

3. Como a Sociolinguística entrou no Brasil?

De modo geral, não fazemos sociolinguística variacionista de orientação laboviana no Brasil. Com poucas exceções, fazemos uma sociolinguística de modo muito peculiar, a sociolinguística variacionista do Brasil (cf. FREITAG, 2016). Nesse modelo, para muitos pesquisadores, a sociolinguística é vista como uma metodologia de geração de dados para testagem de teorias e descrição de língua (ou de variedades). Tanto que "casamentos",

como a Sociolinguística Paramétrica e o Sociofuncionalismo, por exemplo, foram celebrados no Brasil, por pesquisadores brasileiros, tendo pouca repercussão no exterior. De maneira geral, os títulos dos trabalhos assumidos (e reconhecido pelos pares) como sociolinguísticos, como artigos, dissertações e teses, priorizam na posição de núcleo o fenômeno linguístico sob análise, deixando na posição de adjunto a comunidade de fala (referida no nível geográfico) ou algum aspecto social. Vai ser difícil encontrar, dentre os pesquisadores brasileiros da área, um que bata a marca de William Labov, que, durante toda a sua carreira, realizou mais de 1000 entrevistas sociolinguísticas, pessoalmente (LABOV, 2016). O contato com o processo de campo amplia as possibilidades de análise do pesquisador, na medida em que permite captar nuanças que não estão presentes na transcrição (modulação e intensidade da voz) ou mesmo no áudio, como as expressões faciais e o contexto físico da entrevista, por exemplo.

Por outro lado, deter-se exclusivamente na análise de uma grande coleção de dados, já sistematizados e transcritos, permite ao pesquisador aprofundar-se em descrições mais minuciosas das regularidades e no cotejamento de dados a teorias. Com esse trabalho, foi possível depreender uma gramática brasileira, uma norma linguística brasileira calcada na regularidade de usos reais, em dados empíricos. O pesquisador define que isso ou aquilo é português brasileiro a partir do recorte de um fenômeno e da observação da regularidade de seu comportamento em uma amostra linguística sistemática – o banco de dados – que é socialmente estratificada. É dessa abordagem que decorrem generalizações do tipo: maior escolaridade leva à maior presença de marcas de concordância, menor faixa etária leva a maior uso de a gente, e assim por diante.

Mas, para chegar a essas generalizações, houve todo um trabalho anterior de delineamento, coleta e sistematização das amostras linguísticas, que tradicionalmente aparece nas metodologias na voz passiva. Foram selecionados informantes, entrevistas foram realizadas, os áudios foram transcritos. Não está explícito o sujeito agente de todas essas ações. Quem fez a seleção dos informantes? quem gravou as entrevistas? quem transcreveu o áudio? A estratégia da voz passiva aponta para a ideia de empreendimento coletivo que é um banco de dados (sócio)linguísticos. O uso da voz passiva dá a ideia de que não foi uma pessoa, um único pesquisador, quem o fez; foi uma equipe de pesquisadores, com diferentes funções e em diferentes fases. A despessoalização das ações de constituição de um banco de dados leva a questões de autoria e de posse e propriedade, um outro ponto

que, dentro das boas práticas de pesquisa, precisa ser discutido na formação do linguista (quem são os donos dos dados, mais à frente).

Um fator que pode ter contribuído para o formato da sociolinguística no Brasil hoje é o financiamento das pesquisas, que tende a priorizar as propostas coletivas e amplas. Nos retrospectos do GT de Sociolinguística da Anpoll, os coordenadores relatam os esforços dos pesquisadores para a captação de recursos junto a agências de fomento, como a Finep, Capes, CNPq e fundações estaduais, que priorizam propostas coletivas, tanto quanto a pesquisadores, como quanto a áreas do conhecimento envolvidas (talvez a exceção sejam as bolsas incentivo individual – produtividade, jovens pesquisadores, etc.).[5] Nesta conjuntura, projetos que atendam a diferentes objetos de investigação alinhados a uma proposta agrupadora, como a constituição de um banco de dados sociolinguísticos de uma determinada comunidade, tendem a ser priorizados em função da possibilidade de otimização de recursos e esforços.

4. Os donos dos dados

Coletar dados dá trabalho, e não dá para resumir a uma linha em voz passiva todo esse trabalho (FREITAG; MARTINS; TAVARES, 2012). Por isso, é importante discutir a autoria de um banco de dados. São poucos os pesquisadores que podem dizer que têm um banco de dados para chamar de seu propriamente. Isso significa que o pesquisador, único e solitário, desenhou a amostra, foi a campo identificar os falantes que preenchiam o perfil da estratificação da amostra, ele mesmo entrevistou, depois transcreveu e revisou a gravação de áudio de cada um dos falantes que foi entrevistado, para então, enfim, realizar seu trabalho de pesquisa. Há ainda muitos pesquisadores que trabalham assim. Ninguém conhece melhor os seus dados do que este tipo de pesquisador, para o bem e para o mal. No lado positivo, o pesquisador que acompanhou do desenho da pesquisa à transcrição dos dados é capaz de identificar muitos aspectos que não foram captados pela gravação, elucidando dados truncados, por exemplo. No lado negativo, o pesquisador pode estar tão envolvido com seu fenômeno que pode começar a ver dados onde não existe, uma forçada inconsciente na transcrição (ouvir demais). Afinal, os dados são seus e só seus (será?).

[5] Periodicamente, as ações do GT de Sociolinguística são compiladas em coletâneas organizadas, assim como balanços da área, que podem ser conferidos em Brandão (1994), Vandresen (2003), Savedra (2010).

A realidade da maioria dos pesquisadores, no entanto, é diferente. Os bancos de dados costumam ser construídos coletivamente, seguindo mais ou menos uma certa estrutura: um pesquisador ou um grupo de pesquisadores, todos sêniores, desenham uma amostra, definindo as categorias de estratificação dos informantes e o número mínimo de falantes por célula.[6] A partir deste desenho, os pesquisadores iniciantes e em formação trabalham para encontrar os falantes que preencham as células, entrevistá-los; transcrever o áudio da gravação costuma ser tarefa dos pesquisadores iniciantes (iniciação científica), revisar a transcrição costuma ser tarefa dos pesquisadores em formação (mestrandos e doutorandos). E assim um banco de dados não tem um autor; tem vários. Qual a parcela de contribuição de cada um neste processo de constituição de um banco de dados? A entrevista final, gravada e transcrita, só existiria porque um grupo de pesquisadores sêniores delineou a proposta do banco de dados; por outro lado, a concretização do banco de dados projetado só se realizaria em um empreendimento coletivo com os pesquisadores iniciantes e em formação.

Um banco de dados coletivo é um empreendimento colaborativo ou voluntário? O que alguém ganha ao constituir um banco de dados? Como destaquei anteriormente, a sociolinguística brasileira – assim como as demais áreas da ciência – é financiada por agências que valorizam projetos que tenham abrangência amplia, visando o compartilhamento, daí a ênfase em bancos de dados sociolinguísticos, que podem ser utilizados mais de uma vez, para estudar diferentes fenômenos. O financiamento de projetos que visam a constituição de bancos de dados sociolinguísticos recobre as despesas de capital e custeio, mas não de recursos humanos exclusivamente para esse fim; para isso, pesquisadores iniciantes e em formação aderem ao projeto (com bolsa ou não), colaborando com a identificação de um fenômeno linguístico a ser tratado ao tempo em que contribui para a constituição do banco de dados. Seja num plano de trabalho de iniciação científica ou num projeto de mestrado ou doutorado, as regras tácitas estabelecidas entre o grupo que atua na constituição dos bancos de dados é que a contrapartida ao uso de um todo (o banco de dados) é colaborar com uma parte (entrevistas, transcrições, revisões). O acesso ao todo não é gratuito; assim como a colaboração nas partes não é voluntária. E, depois de pronto,

[6] "Falante", "informante", "sujeito", "participante"; depois do termo "dado", estes são outros termos usados como variantes, mas que carregam diferentes significados subjacentes e que mereceriam uma reflexão mais aprofundada (talvez, em outra oportunidade).

um banco de dados, ainda que financiado com o dinheiro público, não é de domínio público: houve um trabalho intelectual por detrás de sua elaboração, desde a concepção até a consecução, que configura autoria. Logo, seu uso deve ser referido apropriadamente.

Se o banco de dados foi financiado com dinheiro público deveria ser público, no sentido de ser compartilhado abertamente, certo? Nem sempre, e não por questões de posse e propriedade, mas por questões de ética em pesquisa. Há bancos de dados que disponibilizam suas amostras na internet. É o caso do Iboruna (http://www.iboruna.ibilce.unesp.br), que exige cadastro para acesso, do SP2010 (http://projetosp2010.fflch.usp.br/node/3) totalmente aberto, do NURC-RE (OLIVEIRA Jr., 2016) e, parcialmente, do PEUL (http://www.letras.ufrj.br/peul/amostras%201.html) e do NURC-RJ (http://www.letras.ufrj.br/nurc-rj/), que disponibilizam as transcrições.

Por outro lado, nem sempre a abertura pública dos dados é possível. A documentação sociolinguística – conjunto de procedimentos adotados para a constituição de bancos de dados sociolinguísticos (cf. FREITAG, a sair A) – não é ingênua nem neutra, menos ainda asséptica, na voz passiva e na terceira pessoa. O processo de seleção dos falantes e todo o período de tempo em que documentador e falante convivem para gerar os dados influenciam nos resultados. Há riscos na documentação sociolinguística que podem e precisam ser minimizados, com a elaboração prévia do desenho da pesquisa, com o treinamento adequado do pesquisador de campo e com o cuidado na divulgação dos resultados.

A regulamentação das questões éticas em pesquisa envolvendo seres humanos (os nossos falantes são seres humanos) parte do ponto de vista que toda a atividade de pesquisa apresenta riscos. Considerando que existem riscos em pesquisa, a submissão de projetos aos comitês de ética tem se tornado um procedimento a cada dia mais constante, em função dos imperativos institucionais, relacionados a agências de financiamento ou ao vínculo do trabalho com a instituição de pesquisa.

Nunca pensamos em fazer mal ao informante por realizar uma entrevista sociolinguística, mas este procedimento sistematicamente adotado nas documentações sociolinguísticas no Brasil, em função de seu roteiro, pode levar o falante a rememorar experiências que podem gerar situações de melancolia. Ao falar de certas passagens de sua vida, um falante pode até chorar. Será que o falante vai concordar que seu choro seja compartilhado publicamente (disponível na internet)? Ou, então, no calor do momento, o falante

faz uma colocação pouco politicamente correta (eu conheço uma entrevista sociolinguística em que o falante expõe sua posição em relação ao aumento de "homossexualismo" nos dias de hoje, culpa do uso de supositórios nas crianças, que se "acostumaram" e que, por isso, os filhos dele nunca usaram supositório). Ainda que o banco de dados anonimize o falante, certas pistas deixadas no decorrer da entrevista podem identificá-lo, como, por exemplo, um falante anonimizado no banco de dados, mas que diz durante a entrevista que era vereador e que trabalhou na instituição tal na época tal. O termo de consentimento livre e esclarecido, requisito obrigatório em bancos de dados cujo projeto de constituição foi apreciado por comitê de ética em pesquisa, resguarda o pesquisador e a instituição que abriga o banco de dados, mas nem sempre o falante é esclarecido ao ponto de ter consciência de que parte de sua vida vai se tornar pública e compartilhada, inclusive na internet.

"Não dê ligância disso" é um dado muito interessante com o verbo *dar*. Mas está numa entrevista sociolinguística (com termo de consentimento livre e esclarecido assinado) em que, em dado momento, a informante vangloria-se do fato de seu pai ter matado um vizinho e ter se evadido para outro estado. Por mais interessante que seja o dado, o crivo final do pesquisador e sua responsabilidade perante os dados e a sua divulgação pode levar a exclusão desta entrevista de um banco de dados, para preservar-se (à informante e a si mesmo).

Anonimato é uma condição, mas não é uma obrigatoriedade. Muitas vezes, não tem como anonimizar; outras vezes, o próprio falante quer ser identificado. Por isso, uma discussão sobre os impactos dessas pequenas decisões é crucial no desenho de uma coleta de dados

Ainda em relação ao desenho do projeto de documentação sociolinguística e as questões éticas subjacentes, é preciso deixar claro ao informante em que célula social ele será alocado. Ao analisar o tratamento do sexo/gênero na sociolinguística brasileira, constatei que, nas perguntas de checagem da entrevista sociolinguística, a estratificação por faixa etária e por escolarização eram conferidas, mas a de sexo/gênero, não (FREITAG, 2015). Ao buscar o falante no campo, para preencher a célula social, presumimos seu sexo/gênero, e essa talvez seja uma pergunta mais simples de se fazer. Mas em estratificações do tipo português culto/português popular, será que é possível pedir ao informante a autoindicação? Será que essa autoindicação bate com aquela que é presumida

pelo documentador, pesquisador de campo? Você, leitor, já se colocou na situação de se dispor a conceder uma entrevista sociolinguística e tempos depois se reconhecer como um falante de português popular?

O treinamento de campo do documentador é uma parte importante do processo de constituição de um banco de dados, mas que, muitas vezes é aligeirada ou simplesmente suprimida. O diferente causa estranhamento. O documentador não pode rir nem corrigir o falante, só porque ele falou "apesar dos apesares" ou "nóis vinhemo". Ainda que seja uma reação inconsciente, traz constrangimento ao falante, é um risco que pode ser evitado. Assim como podem ser evitados os desgastes decorrentes de uma transcrição ortográfica adaptada (aquela que marca traços da fala).

Considerar as questões envolvendo ética em pesquisa na constituição de um banco de dados é uma necessidade relativamente recente, decorrente das imposições de agências de fomento, que, de certa forma, são as mesmas imposições (a necessidade de financiamento) que deram a cara da Sociolinguística do Brasil, com bancos de dados coletivos para a descrição linguística. No entanto, sempre houve *outliers*, como em todos em campos da ciência, e, nos últimos anos, trabalhos que tomam como ponto de partida a comunidade e seus fatores constitutivos, e não um fenômeno linguístico – como a variação e o efeito das redes no Morro dos Caboclos (FERRARI, 1994) e a caracterização sociolinguística da comunidade de fala de São Paulo (OUSHIRO, 2015), por exemplo – têm emergido, sinalizando uma mudança de tendência, que traz novos desafios e novas discussões teórico-metodológicas na sociolinguística brasileira. Se a ciência brasileira é financiada pelos pais (http://revistagalileu.globo.com/Revista/noticia/2016/07/ciencia-no-brasil-e-bancada-pelos-pais.html), talvez este seja o momento de mudar um ciclo.

5. O quão dados são os dados de um banco de dados?

A abordagem do dado/doado costuma ser associada ao acaso, aleatório, no sentido que o dador/doador seria uma pessoa qualquer. Tanto que os estudos sociolinguísticos, pelo menos no Brasil, costumam assumir que suas amostras são probabilísticas, aleatórias e estratificadas, o que quer dizer que qualquer falante da população delimitada teria o mesmo número de chances de ser escolhido para preencher a cota de cada estrato. Na prática, sabemos que a teoria é outra: os falantes que colaboram para a constituição de um banco de dados sociolinguística são aqueles que dispõem de tempo para colaborar,

cedendo no mínimo uma hora de sua vida para essa finalidade. Imperam a disponibilidade e a voluntariedade, o que leva a um modelo de cotas por conveniência e julgamento. A conveniência permite a operacionalidade da coleta, no entanto, 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. E, por serem pautadas na conveniência, limitam a replicabilidade, na medida que há um viés de seleção (FREITAG, a sair B).

O viés de seleção pode se dar no momento da transcrição impressionística dos dados: aquilo que é transcrito nem sempre é o que o falante falou, mas o que o transcritor ouviu: a depender do seu fenômeno, haverá maior sensibilidade para percebê-lo e marcá-lo, em transcrições ortográficas adaptadas ou fonéticas; ou, então, a variante do transcritor se sobrepõe à variante realizada pelo falante, num processo inconsciente de acomodação linguística. Por isso, o dado dado não é tão aleatório assim, nem para um lado, nem para o outro. Sempre haverá um viés de seleção.

Considerações

Ter a oportunidade de discussão de aspectos que permeiam a formação do linguista é importante para formarmos uma cultura de crítica e avaliação do fazer da nossa área, como um todo, ao tempo que podemos também nos posicionar e discutir, resgatando uma cultura de debate (volto a insistir nos debates da Delta).

Em tempo: recorrendo ao oráculo do século XXI, não achei nenhuma ocorrência de "dadidão", mas 1090 ocorrências para "dadidade", basicamente relacionadas ao campo da ontologia (consulta ao Google, em páginas em português, na data limite da chamada para a publicação desta edição).[7] Embora "dadidade" seja mais produtiva, acho que "dadidão" fica melhor. Linguistas, eis um tema para novas pesquisas!

[7] Na revisão, apareceram 1750 ocorrências para "dadidade" e ainda nenhuma para "dadidão".

Agradecimentos

Agradeço à minha orientadora, Edair Maria Görski, que me permitiu entrar num grupo de pesquisa e fazer o todos faziam, e também por me permitir quebrar as regras de vez em quando. Agradeço aos pareceristas, a um, por suas contribuições para novas discussões e, ao outro, por me lembrar do *habitus* da academia. E agradeço à Livia Oushiro, por sua leitura atenta e sincera.

REFERÊNCIAS

BOURDIEU, Pierre. *A economia das trocas simbólicas*. Trad. Sérgio Miceli. São Paulo: Perspectiva, 1974.

BRANDÃO, Silvia Figueredo. GT de Sociolinguística. *Revista da Anpoll*, v. 1, n. 1, 1994.

DILLINGER, Mike. Forma e função na lingüística. *DELTA*, v. 7, n. 1, p. 395-407, 1991.

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

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: Blücher, 2015, p. 17-74.

FREITAG, Raquel Meister Ko. Amostras sociolinguísticas: probabilísticas ou por conveniência? (a sair A).

FREITAG, Raquel Meister Ko. *Documentação sociolinguística – coleta de dados e ética em pesquisa*. São Cristóvão: EdUFS (a sair B).

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

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**, v. 56, n. 3, p. 917-944, 2012.

KATO, Mary A. Formas de Funcionalismo na Sintaxe. *DELTA*, v. 14, n. spe, p. 00,1998.

LABOV, William. Afterword: Where are we now? *Journal of Sociolinguistics*, v. 20, n. 4, p. 581–602, 2016.

LOBATO, Monteiro. *Reinações de Narizinho*. Rio de Janeiro: Globo Livros, 2014 [1933].

NASCIMENTO, M. do. Teoria gramatical e mecanismos funcionais do uso da língua. *Delta*, v. 6, n. 1, p. 83-98, 1990.

OLIVIERA Jr, Miguel. NURC Digital Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do Projeto da Norma Urbana Linguística Culta (NURC). *CHIMERA: Romance Corpora and Linguistic Studies*, v. 3, n. 2, p. 149-174, 2016.

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

POSSENTI, Sirio. O dado dado e o dado dado (o dado em análise do discurso). In: POSSENTI, Sírio. *Os limites do discurso: ensaios sobre discurso e sujeito*. São Paulo: Parábola, 2009, p. 23-31.

SAVEDRA, Mônica Maria Guimarães. Estudos e pesquisas em sociolinguística no contexto plurilíngue do Brasil. *Revista da Anpoll*, v. 1, n. 29, 2010.

SCHER, Ana Paula. Nominalizações em-ada em construções com o verbo leve dar em português brasileiro. *Letras de Hoje*, v. 41, n. 1, 2006.

SILVA, Leilane Ramos da. Construções lexicais complexas com o verbo dar: estruturas de significados ou instrumentos de construção de sentidos. *Estudos Lingüísticos*, v. 34, p. 563-568, 2005.

VANDRESEN, Paulino. A trajetória do GT de Sociolingüística da ANPOLL–1985-2001. In: C. Roncarati; J. Abraçado (Org.). *Português Brasileiro-Contato Lingüístico e História***.** Rio de Janeiro: 7Letras, 2003, p. 13-29.

VOTRE, Sebastião Josué; NARO, Anthony Julius. Mecanismos funcionais do uso da língua. *DELTA*, v. 5, n. 2, p. 169-184, 1989.

ABSTRACT: The goal of this text is to raise a few remarks on the entity “linguistic data” in order to contribute to the formation of linguists. The organization of linguistic datasets considering their authorship, possession, ownership and randomness are the aspects under discussion.

KEYWORDS: Linguistic data. Authorship. Randomness.

FREITAG, Raquel M.K. A dadidade (ou dadidão) do dado, *Linguística Rio*, vol.3, n.1, maio de 2017.

ISSN: 2358-6826

Enviado: 19 de fevereiro de 2017
Aceito: 18 de abril de 2017
Online: 02 de junho de 2017