



## COLETA DE DADOS PARA PESQUISAS SOCIOLINGÜÍSTICAS (EM TEMPO DE PANDEMIA)

Livia OUSHIRO<sup>1</sup>

### Resumo

Este trabalho tem como objetivos (i) reportar a experiência e as reflexões da disciplina LL051–Metodologia da Investigação Sociolinguística do Programa de Pós-Graduação em Linguística do IEL/UNICAMP, no primeiro semestre de 2020, cujo tópico principal foi o desenvolvimento de trabalho de campo; e (ii) refletir criticamente sobre a coleta de dados na Sociolinguística. A disciplina consistiu em discussões, planejamentos e coletas de dados através de elicitación (CHELLIAH, 2014), questionários (CAMPBELL-KIBLER, 2013), monitoramento midiático (D'ARVY; YOUNG, 2012) e entrevistas sociolinguísticas (BECKER, 2013), tendo em mente o *continuum* naturalidade-controle (NAGY, 2006) e considerações éticas para com os participantes (ECKERT, 2013). Argumenta-se que o contexto de comunicação digital deve conduzir o sociolinguista à análise de novos contextos da língua em uso, mas também à revisão dos procedimentos éticos de suas coletas.

**Palavras-chave:** trabalho de campo; ética em pesquisa; elicitación; questionários; monitoramento midiático.

### Introdução

Este trabalho<sup>2</sup> é fruto de uma reflexão coletiva feita junto aos alunos de graduação e de pós-graduação da Unicamp. Um dos objetivos é o de relatar as experiências na disciplina LL051–Metodologia da Investigação Sociolinguística, oferecida no primeiro semestre de 2020 no Programa de Pós-Graduação em Linguística do IEL-UNICAMP, durante os primeiros meses de isolamento social em decorrência da pandemia de COVID-19. O objetivo do curso foi tanto teórico – ou seja, a reflexão crítica sobre o que acarreta coletar dados de uma ou outra maneira –, quanto prático – o desenvolvimento efetivo de coletas de dados por

<sup>1</sup> Universidade Estadual de Campinas. E-mail: [oushiro@unicamp.br](mailto:oushiro@unicamp.br).

<sup>2</sup> Uma versão deste trabalho também foi apresentada no X Encontro de Sociolinguística em 3.dez.2020. Disponível em <https://youtu.be/zIWHt0ElcAg>. Último acesso em 20.dez.2020. Agradeço a todos os participantes do curso LL051 pelas reflexões e pelo companheirismo durante o primeiro semestre de isolamento social: Abdulai Danfá, Aline Jéssica Pires, Bruna Barbosa Louzada, Isabel P. de Lima e Souza, Gustavo C. Pinheiro, Joyce Mattos, Julia Adams, Karin Vivanco, Leila Maria Tesch, Lucas Pereira Eberle, Raíssa Silva Santana, Tereza M. Barboza, Thuany T. de Figueiredo e Vitória B. H. Lisboa.

parte dos alunos. Neste artigo, o principal objetivo é tratar dos cuidados éticos, analíticos e práticos ao coletar dados de outras fontes que não a entrevista sociolinguística presencial.

A Unicamp foi a primeira universidade no Brasil a suspender as aulas, a partir de 13 de março de 2020. Para a disciplina LL051, especificamente, os alunos e a docente começaram a refletir sobre como os conteúdos e discussões poderiam ser adaptados para o formato *online*: deixar toda a parte prática de lado e nos concentrarmos apenas nas discussões teóricas? Aguardar o fim da pandemia e para posterior retomada? Como fazer trabalho de campo durante o isolamento social? Esta última pergunta, que surgiu por conta da disciplina, certamente também foi feita por muitos outros pesquisadores e pós-graduandos, à medida que a pandemia se estendeu para muito além do que se havia imaginado.

A resposta foi encontrada em Miller (2017), especialista em Antropologia Digital e etnografia *online*,<sup>3</sup> que afirma: “não importa o que os povos do mundo fazem, é isso que queremos estudar (...) se o mundo se torna digital, nós nos tornamos digitais.” Essa afirmação se aplica perfeitamente aos sociolinguistas, que estudam a língua em seu contexto social: se as pessoas estão se comunicando pela Internet, também devemos analisar essas interações, e também podemos coletar dados *online*.

Na prática, nós sociolinguistas já fazemos isso: existem trabalhos que se baseiam em dados de Twitter, Facebook, Instagram, WhatsApp ou canais do YouTube. Mas a Sociolinguística Variacionista mais tradicional costuma lidar com textos orais e com textos comuns, normalmente gravados em entrevistas sociolinguísticas. Existem dois tipos de reação a trabalhar com dados de outras fontes, como aqueles advindos de textos escritos ou da mídia: uma reação é a de resistência, já que esses dados não seriam, em princípio, tão espontâneos ou representativos da fala cotidiana; outra reação possível é a de não entrever problema algum, simplesmente transpondo a metodologia de extração e análise de dados de entrevistas para outros *corpora*. Aqui se argumenta que outras fontes de dados, para além da entrevista, podem fornecer rico material para estudo da língua em uso; por outro lado, a transposição de fontes não é sem consequências.

Labov (1972), em um artigo metodológico, avalia diferentes fontes de dados e as particularidades de cada um:

Current difficulties in achieving intersubjective agreement in linguistics require attention to principles of methodology which consider sources of error and ways to eliminate them. The methodological assumptions and practices of various branches of linguistics are considered from the standpoint of the types of data gathered: texts, elicitations, intuitions and

<sup>3</sup> *Digital Anthropology* Daniel Miller. Disponível em: [https://www.youtube.com/watch?v=XNus-xZ7\\_6Y&t=55s](https://www.youtube.com/watch?v=XNus-xZ7_6Y&t=55s); Como conduzir uma etnografia durante o isolamento social. Disponível em: <https://www.youtube.com/watch?v=NSiTrYB-0so&t=737s>. Último acesso em 20 dez. 2020. Agradeço a Thuany Figueiredo pela indicação desses vídeos.

observations. Observations of the vernacular provide the most systematic basis for linguistic theory, but have been the most difficult kinds of data for linguists to obtain; techniques for solving the problems encountered are outlined. Intersubjective agreement is best reached by convergence of several kinds of data with complementary sources of error. (LABOV, 1972, p. 97)

O primeiro ponto a se destacar dessa citação é que todo tipo de coleta contém fontes de erro, o que deve conduzir o pesquisador a identificá-las e eliminá-las, a fim de atingir a concordância intersubjetiva em Linguística. Labov considera diferentes tipos de dados: textos, eliciações, intuições e observações do vernáculo, sendo este último tipo a base mais sistemática para a teoria linguística, mas também o tipo de dados mais difícil de se obter. A conclusão geral de Labov é que a concordância intersubjetiva será atingida “pela convergência de diversos tipos de dados, com fontes complementares de erro.” O ideal, portanto, é a combinação de vários tipos de coleta, pois nas fontes em que um tipo de dados é limitado, outro pode complementá-lo.

Para essa formulação de Labov (1972), é central a ideia de *vernáculo*, que o autor define como “o modo como as pessoas falam quando não estão prestando atenção à própria fala” e como “a variedade adquirida na infância”. Labov considera o vernáculo a forma mais sistemática e regular da língua, o que seria o verdadeiro objeto do linguista. Entretanto, Labov (2006 [1966]) reconhece que a entrevista sociolinguística gera o Paradoxo do Observador: queremos observar as pessoas do modo como falam em seu dia-a-dia, mas ao colocá-las em situação de gravação, automaticamente já não se está mais observando a fala espontânea. Nesse sentido, o objetivo da entrevista sociolinguística é o de obter uma gama de estilos de fala do mesmo indivíduo dentro de uma mesma situação controlada de coleta de dados.

*Naturalidade e controle* são os dois extremos que regem a coleta de dados. Nagy (2006) esquematiza um *continuum* entre esses polos, citando vários tipos de coletas (Figura 1). No extremo mais abaixo, estão os dados elicitados, que são mais fáceis de coletar e de analisar por haver maior controle do pesquisador. Subindo este *continuum*, tem-se a produção de sons isoladas, a repetição de palavras, a leitura de uma lista de palavras, conjugações, traduções, sentenças construídas, a descrição de figuras, narrativas, conversas com um pesquisador (o que equivale à entrevista sociolinguística) e conversas com um membro do mesmo grupo. Neste outro extremo, tem-se os dados representativos da fala natural, necessariamente com menor controle do pesquisador.

representativo da fala natural

▪

### **Natural**

conversa com membro do grupo  
conversa com pesquisador  
narrativas  
descrição de figuras  
sentenças construídas  
traduções  
paradigmas, p.ex. conjugações  
lista de palavras  
repetição de palavras  
produção de sons isolados

### **Elicitado**

*fácil de coletar e de analisar*

**Figura 1** – *Continuum* de técnicas de pesquisa sociolinguística<sup>4</sup>.

Considerando os tipos de dados com que se decidiu trabalhar no curso de pós-graduação LL051, pode-se considerar que a elicitación de dados de uma língua desconhecida é o menos espontâneo e mais controlado, seguido dos questionários e das entrevistas sociolinguísticas, sendo o monitoramento midiático o tipo de dados mais espontâneos pois, mesmo que dados de mídia possam ser previamente planejados e editados, eles não foram produzidos para análise linguística, mas para outros fins. O único tipo de coleta que não foi possível implementar durante o curso foi o levantamento rápido e anônimo.<sup>5</sup>

## **As coletas**

### **Elicitación (e Entrevistas Sociolinguísticas)**

A primeira coleta desenvolvida foi a de elicitación de dados linguísticos de uma língua desconhecida pelo aluno<sup>6</sup> (CHELLIAH, 2014), cujas considerações também se aplicam em grande medida para a gravação de entrevistas sociolinguísticas (BECKER, 2013). Não houve grandes problemas técnicos para essa coleta. Os alunos agendaram encontros com os participantes via Google Meet ou plataformas similares, enviaram o TCLE por e-mail e obtiveram consentimento dos participantes oralmente quando estes não podiam imprimir ou assinar o documento digitalmente. A maioria dos alunos gravou a interação com áudio e vídeo, o que abre a possibilidade de análises não apenas da fala, mas também de análises multimodais. Embora alguns alunos tenham reportado problemas no agendamento com o

<sup>4</sup> Traduzido de Nagy (2006, p. 390).

<sup>5</sup> Dentro do esquema de Nagy (2006), é interessante tentar encaixar os levantamentos rápidos e anônimos, já que esse tipo de levantamento envolve amplo controle do pesquisador, são dados fáceis de coletar e de analisar, e ao mesmo tempo são maximamente espontâneos. Trata-se de um tipo de coleta subutilizado na Sociolinguística, pois ainda que se obtenham dados limitados, eles têm a vantagem do controle e da espontaneidade.

<sup>6</sup> Esses dados incluíam uma lista de 40 itens lexicais, o sistema de pluralização ou de marcação temporal em verbos daquela língua e a estrutura de orações relativas.

participante, que desmarcou o encontro algumas vezes, a maioria não teve problemas nesse sentido; aparentemente, as pessoas se mostram mais disponíveis para conversas *online* do que em encontros presenciais – algo ainda a ser mais bem avaliado em experiências futuras e em comunidades de perfis diversos.

Se isso for verdadeiro, deve-se considerar se mesmo depois da pandemia não deveríamos incorporar a coleta digital definitivamente. Por outro lado, a coleta exclusivamente pela Internet pode acabar restringindo o perfil dos participantes de nossas pesquisas, daqueles que têm maior dificuldade de acesso à Internet ou de manipulação de recursos digitais, como populações mais isoladas e/ou idosas.

### **Questionários**

Os questionários (CAMPBELL-KIBLER, 2013) são um tipo de coleta que já se tem sido realizado amplamente via Internet, mesmo antes da pandemia. Aqui, no entanto, vale ressaltar que pode haver diferenças nas respostas dos participantes a depender de se a coleta for presencial ou *online* – o que, aliás, também se aplica a entrevistas sociolinguísticas. Em um estudo de percepções sobre a pronúncia de /r/ em coda como tepe ou retroflexo (OUSHIRO, 2015, 2019a), observaram-se diferenças quantitativas, mas não qualitativas, entre as respostas coletadas presencial ou virtualmente: apesar de os julgamentos serem na mesma direção, os participantes deram respostas relativamente mais extremas nas respostas *online*. Isso, na verdade, é um ponto positivo para a coleta remota: os participantes se sentem menos inibidos em emitir julgamentos quando não estão na presença do pesquisador.

Para essa coleta, a maior dificuldade dos alunos não foi na coleta em si, mas no planejamento prévio, para definir claramente quais eram as perguntas de pesquisa e qual era a tarefa que o participante deveria executar. Outro ponto preocupante foi o fato de que a grande maioria decidiu usar o Google Forms por ser mais fácil e intuitivo, mesmo que a plataforma não ofereça muitas opções de tipos de perguntas. É necessário estar alerta ao fato de que muitas vezes se restringem as questões de pesquisa por conta das plataformas; cabe ao pesquisador selecionar as ferramentas que permitam responder mais propriamente suas questões de pesquisa (OUSHIRO, 2019b).

### **Monitoramento midiático**

O *monitoramento midiático* é aqui entendido como a análise de textos orais ou escritos, tanto da mídia tradicional (TV, rádio) quanto de mídias mais recentes (Facebook, Twitter, Instagram etc., ou de plataformas de interação online como blogs e MMOs<sup>7</sup>), que não

---

<sup>7</sup> Sigla em inglês para *massive multiplayer online game* (jogo *online* com número massivo de jogadores).

foram produzidos originalmente para análise linguística, diferentemente de entrevistas sociolinguísticas, eliciações e questionários. Com o advento e a difusão da Internet, sobretudo a partir da década de 2000, grandes quantidades de dados estão disponíveis na rede (ANDROUTSOPOULOS, 2013), a maior parte dos quais mais espontâneos do que se o próprio pesquisador elicitá-los de algum modo.

Buzato, D'Angelis e Motta (2017), contudo, advertem que nem todo dado que está na Internet pode ser coletado. No caso específico de blogs, sites pessoais, perfis de redes sociais, WhatsApp e canais de YouTube, não há diretriz definitiva: é necessário olhar caso a caso. Devem ser consideradas tanto questões jurídicas, como de *copyright* e direitos de uso, mas também – e, pode-se dizer, principalmente – a responsabilidade do pesquisador para com quem produziu esses dados. Nisso se incluem tanto o direito à privacidade quanto a garantia de que não se corra o risco de haver qualquer prejuízo ao participante (ECKERT, 2013).

D'Arcy e Young (2012), em artigo sobre coleta de dados em mídias digitais, incitam os pesquisadores a se questionar sobre qual é o seu papel na coleta de dados *online*, ao retomar a tipologia proposta por Bell (1984) em sua Teoria de Design da Audiência. Bell (1984) classifica a audiência entre destinatários, ouvintes, ouvintes casuais e bisbilhoteiros, que se definem a depender de se é a pessoa a quem se dirige; se aquela pessoa é ratificada, i.e., tem permissão para ouvir aquela conversa; e se aquela pessoa é conhecida pelos falantes, i.e., sabe-se que ela é um potencial ouvinte da interação. Na coleta *online*, muitas vezes o pesquisador acaba agindo como ouvinte casual (conhecido mas não ratificado) ou como bisbilhoteiro (não conhecido e não ratificado pelo produtor do dado) – e isso tem sérias implicações éticas.

A pesquisa de Adams (2019)<sup>8</sup> é um exemplo de como lidar com essas questões de modo ético. Sua análise sobre *preposition stranding* (“aquele assunto que temos que falar sobre”) e *orphaning* (“vamos conversar sobre?”) no Português Brasileiro, com dados do Twitter, compreende um *corpus* com mais de 10 milhões de palavras. Para essa quantidade de *tweets*, seria quase impossível obter o consentimento de todos os produtores dos dados. A pesquisadora, no entanto, avalia que o Termo de Uso da plataforma não corresponde a um TCLE, e que mesmo se o usuário tenha ciência de que seus dados podem ser acessados por qualquer pessoa, isso não é sinônimo de ter dado consentimento para participar de uma pesquisa. Tendo em vista a impossibilidade de obtenção do TCLE, mas também sua responsabilidade como pesquisadora para não trazer potenciais prejuízos aos produtores dos dados, os procedimentos consistem em não só anonimizar o usuário, mas também em trocar itens lexicais não essenciais por sinônimos e nunca citar dados polêmicos (como

<sup>8</sup> CAAE: 07497019.1.0000.8142.

manifestações abertas de racismo ou declarações de intolerância religiosa, política etc.), a fim de dificultar e não incitar a possibilidade de busca reversa na Internet.

Em outros casos, contudo, é impossível anonimizar o produtor dos dados, como quando se utilizam dados de canais do YouTube. Lucca (2017), por exemplo, analisou a variação estilística na fala de um motoboy de São Paulo em seu canal Motoka Cachorro,<sup>9</sup> para a qual entrou em contato com o produtor do canal a fim de obter seu consentimento para uso dos dados. Esse procedimento é indispensável.

Vale dizer que há canais de todo tipo no YouTube; muitos deles têm uma abundância de dados, muitas vezes longitudinais, mais espontâneos do que os de entrevistas sociolinguísticas, mesmo passando por planejamento e por edição. Esses dados trazem consigo a dificuldade – mas também a possibilidade – de lidar com textos multimodais: textos escritos, sons, imagens, vídeos inseridos, que ocorrem simultaneamente e que não devem ser ignorados, mas sim incorporados à análise. Tudo isso nos faz ficar mais próximos do vernáculo e das interações do modo como ocorrem no dia-a-dia das pessoas, tal como Labov já havia defendido na década de 1970.

### Considerações finais

“Novas” fontes de dados, como questionários *online*, entrevistas *online* e dados da Internet, abrem muitas perspectivas para estudos sociolinguísticos e podem fornecer dados tão ou mais naturais quanto os de entrevistas sociolinguísticas tradicionais. Todos os tipos de coleta têm suas vantagens e desvantagens, mas a utilização de vários tipos pode compensar as fontes de erro de cada um. As presentes reflexões partiram principalmente da situação que vivemos em 2020, a de isolamento social e da pandemia, que objetivamente nos impede de fazer coleta de dados do modo mais tradicional; tal quadro também nos conduz a reflexões mais gerais sobre o trabalho de campo do sociolinguista, independentemente da pandemia.

A Internet hoje se apresenta como uma enorme fonte de dados, mas esses dados requerem cuidados éticos e a responsabilidade do pesquisador – como requer, aliás, qualquer tipo de coleta. Dentro desse cenário, o sociolinguista tem a oportunidade de analisar novos contextos da língua em uso, mas também precisa revisar os procedimentos éticos de suas coletas. Enquanto aguardamos a volta à “normalidade”, já sabemos de antemão que depois de isso tudo passar, muito não será mais do mesmo jeito. Cabe a nós pensar o que de toda essa experiência vale a pena ser mantido ou não. Já existe uma vasta bibliografia sobre trabalho de campo *online* e ética de pesquisa na Internet, mas ainda não há tantas discussões sobre o assunto no Brasil – algo que precisamos desenvolver urgentemente.

<sup>9</sup> Disponível em: <https://www.youtube.com/user/patocrazymotoboyvlog>. Último acesso em 30 nov. 2020.

## REFERÊNCIAS

- ADAMS, J. B. **Um estudo sobre *preposition stranding* e *orphaning* em falantes de Português Brasileiro**. Pôster apresentado no ABRALIN 50. Ms, 2019.
- ANDROUTSOPOULOS, J. Online data collection. In: MALLINSON, C.; CHILDS, B.; VAN HERK, G. (eds.) **Data collection in Sociolinguistics: methods and applications**. New York and London: Routledge, 2013, p. 236–249.
- BECKER, K. The sociolinguistic interview. In: MALLINSON, C.; CHILDS, B.; VAN HERK, G. (eds.) **Data collection in Sociolinguistics: methods and applications**. New York and London: Routledge, 2013, p. 91–100.
- BELL, A. Language style as audience design. **Language in Society**, vol. 13, p. 145–204, 1984.
- BUZATO, M.; D'ANGELIS, W.; MOTTA, T. **Orientações sobre o Comitê de Ética em Pesquisa**. Comunicação apresentada no IEL/UNICAMP em 6 abr. 2017. Ms, 2017.
- CAMPBELL-KIBLER, K. Language attitude surveys: speaker evaluation studies. In: MALLINSON, C.; CHILDS, B.; VAN HERK, G. (eds.) **Data collection in Sociolinguistics: methods and applications**. New York and London: Routledge, 2013, p. 142–146.
- CHELLIAH, S. Fieldwork for language description. In: PODESVA, R. J.; DEVYANI, S. (eds.) **Research methods in Linguistics**. Cambridge: Cambridge University Press, 2014, p. 51–73.
- D'ARCY, A.; YOUNG, T. M. Ethics and social media: Implications for sociolinguistics in the networked public. **Journal of Sociolinguistics**, vol. 16, n. 4, p. 532–546, 2012.
- ECKERT, P. Ethics in linguistic research. In: PODESVA, R. J.; DEVYANI, S. (eds.), **Research methods in Linguistics**. Cambridge: Cambridge University Press, 2013, p. 11–26.
- LABOV, W. Some principles of linguistic methodology. **Language in Society**, vol. 1, n. 1: p. 97–120, 1972.
- LABOV, W. **The social stratification of English in New York City**. São Paulo: Cambridge University Press, 2006 [1966].
- LUCCA, J. F. **O diário moderno de um motoboy em São Paulo: construção identitária e recursos estilísticos**. Tese (Doutorado em Linguística). Instituto de Estudos da Linguagem, UNICAMP, Campinas-SP, 2017.
- MILLER, D. **Digital Anthropology**. Disponível em: <https://youtu.be/NSiTrYB-0so>. 2017. 1 vídeo (20 min). Último acesso em: 20 dez. 2020.
- NAGY, N. Experimental methods for study of linguistic variation. In: BROWN, K. (ed.). **Encyclopedia of Language & Linguistics**. 2ª ed. vol. 4. Oxford: Elsevier, 2006, p. 390–394.
- OUSHIRO, L. **Identidade na pluralidade: avaliação, produção e percepção linguística na cidade de São Paulo**. Tese (Doutorado em Linguística). FFLCH, USP, São Paulo, 2015.
- OUSHIRO, L. A computational approach for modeling the indexical field. **Revista de Estudos da Linguagem**, vol. 27, n. 4, p. 1737–1786, 2019a.
- OUSHIRO, L. Questões e métodos: vogais médias pretônicas na fala de migrantes nordestinos em situação de contato dialetal. In: VIEIRA, M. S. M.; WIEDEMER, M. L. **Dimensões e experiências em Sociolinguística**. São Paulo: Blucher, 2019b, p. 157–187.