Blucher Open Access

Metodologia de Coleta e Manipulação de Dados

EM SOCIOLINGUÍSTICA

Raquel Meister Ko. Freitag
Organizadora

Blucher

METODOLOGIA DE COLETA E MANIPULAÇÃO DE DADOS EM SOCIOLINGUÍSTICA

Esta obra foi financiada com recursos do edital CAPES/FAPITEC/PROMOB, projeto "Da expressividade da língua ao mal na literatura: bases de pesquisa do PPGL/UFS".

Blucher

Metodologia de Coleta e Manipulação de Dados em Sociolinguística

Editora Edgard Blücher Ltda.

Blucher

Rua Pedroso Alvarenga, 1245, 4º andar 04531-012
São Paulo – SP – Brasil
Tel.: 55 11 3078-5366
contato@blucher.com.br
www.blucher.com.br

Segundo o Novo Acordo Ortográfico, conforme 5ª ed. do *Vocabulário Ortográfico da Língua Portuguesa*, Academia Brasileira de Letras, março de 2009.

FICHA CATALOGRÁFICA

Metodologia de coleta em manipulação de dados em sociolinguística / organizado por Raquel Meister Ko. Freitag. - São Paulo : Blucher, 2014.

ISBN 978-85-8039-085-8

1. Socioliguística 2. Linguagem e línguas – variações 3. Ciências sociais – metodologia 4. Linguística – metodologia I. Freitag, Raquel Meister Ko.

14-0413 CDD 410

Índices para catálogo sistemático:
1. Sociolinguística – aspectos sociais

METODOLOGIA DE COLETA E MANIPULAÇÃO DE DADOS EM SOCIOLINGUÍSTICA

Raquel Meister Ko. Freitag
organizadora

Conteúdo

APRESENTAÇÃO

Raquel Meister Ko. Freitag

Bancos de dados linguísticos de fala – especialmente os elaborados para a pesquisa de orientação sociolinguística variacionista – têm sido fonte privilegiada para a descrição do português brasileiro. Em atendimento às exigências, como as diretrizes de ética em pesquisa envolvendo seres humanos, além dos custos envolvidos e a necessidade de adequação tecnológica para a manipulação de grandes volumes de dados, é preocupação dos pesquisadores da área que haja a definição de procedimentos metodológicos padronizados para a organização de novos bancos de dados em alinhamento com os já constituídos.

Como ação do projeto "Da expressividade da língua ao mal na literatura: bases interinstitucionais de pesquisa do PPGL/UFS" (PROMOB/CAPES/FAPITEC), o *Workshop* "Metodologia de Coleta e Manipulação de Dados em Sociolinguística", realizado nos dias 16, 17 e 18 de fevereiro de 2014, na Universidade Federal de Sergipe, com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Programa de Apoio a Eventos no País (PAEP), da Fundação de Apoio à Pesquisa e à Inovação Tecnológica do Estado de Sergipe (FAPITEC) e o Programa de Auxílio ao Pesquisador para a Realização de Reunião ou Evento Científico e Tecnológico (PRAEV), reuniu pesquisadores e estudantes da área de Sociolinguística com o objetivo de compartilhar, fomentar e cooperar com a discussão do desenvolvimento de protocolos de pesquisa para padronizar os procedimentos de organização de bancos de dados sociolinguísticos.

Resultante do *Worskhop* é esta coletânea de textos dividida em quatro partes. A primeira parte contempla aspectos pré-coleta de dados: a adequação aos preceitos éticos de pesquisa envolvendo seres humanos e a escolha de equipamentos para a gravação da coleta de dados. A segunda parte é dedicada ao compartilhamento de experiências das coletas aos moldes clássicos da Sociolinguística. A terceira parte contempla coletas de dados que considerem aspectos da terceira onda de estudos da Sociolinguística em redes sociais e comunidades de práticas. A quarta e última parte trata de ferramentas computacionais para o armazenamento e manipulação de dados linguísticos.

Todos os textos aqui apresentados têm caráter tutorial, em linguagem e terminologia que seja acessível ao pesquisador iniciante, mas que trazem também discussões prementes aos pesquisadores mais avançados na área. Esperamos colaborar para o aprimoramento da pesquisa sociolinguística brasileira, na medida em que o estabelecimento de padrões pode facilitar a realização de estudos contrastivos entre as diferentes variedades do português, contribuindo, assim, para a identificação e refinamento de generalizações e princípios universais.

CAPÍTULO 1

ASPECTOS LEGAIS ENVOLVIDOS NA COLETA DE DADOS LINGUÍSTICOS

Ricardo Nascimento Abreu

INTRODUÇÃO

O processo de coleta de dados para a constituição dos *corpora*, não só para as pesquisas em linguística, mas para todas as ciências humanas e sociais, está completamente adstrito às mesmas regras éticas que valem para as ciências médicas. Esse sufocamento teve origem após a Segunda Guerra Mundial, dado o estarrecimento da comunidade internacional com as notícias advindas das análises das práticas de "higiene racial" do governo nazista. A decisão de intervir, tutelando os direitos das pessoas submetidas às pesquisas científicas e, principalmente, a necessidade de estabelecer obrigações e responsabilidades para instituições e seus pesquisadores, colocou-se como algo urgente no cenário do pós-guerra. Assim, o Código de Nuremberg (1947), a Declaração Universal dos Direitos Humanos (1948) e a Declaração de Helsinque (1964) surgem, no âmbito do Direito Internacional, como os principais alicerces nos quais se apoiaram as constituições e legislações infraconstitucionais no mundo para fins de regramento ético nas pesquisas envolvendo seres humanos.

No Brasil, a via de entrada desses parâmetros ético-normativos foi o Conselho Nacional de Saúde que, primeiramente através da Resolução n. 196, de 10 de outubro de 1996, e após, com a Resolução n. 466, de 12 de dezembro de 2012 – vem estabelecendo os parâmetros éticos da pesquisa envolvendo seres humanos. O caráter excessivamente biomédico dessas resoluções tem despertado, nos

últimos anos, severas críticas dos pesquisadores das áreas de ciências sociais e humanas, posto que diversos aspectos atinentes às pesquisas dessas áreas não foram contemplados pelas resoluções e, em várias questões, até mesmo foram inviabilizados. Este texto se debruça sobre as principais implicações legais vinculadas às pesquisas nas humanidades, com ênfase na coleta de dados linguísticos em pesquisas que envolvam a participação de seres humanos.

1. ASPECTOS ÉTICOS E SUA REGULAÇÃO NO BRASIL

Apesar de encontrarmos na história das ciências registros de elaborações que apontavam para o sentido de estabelecer limites aos procedimentos experimentais, somente no alvorecer da segunda metade do século XX é que a comunidade internacional passa a tutelar os direitos dos sujeitos submetidos às pesquisas científicas, bem como as obrigações e responsabilidades das instituições e seus pesquisadores.

O término da Segunda Guerra Mundial trouxe consigo, entre outras questões, um sem-número de alertas para a comunidade científica internacional sobre os experimentos envolvendo seres humanos, realizados pelos nazistas, na busca da chamada "higiene racial". Esses experimentos afrontavam um dos princípios do direito moderno que, naquela ocasião, fortalecia-se cada vez mais: o princípio da dignidade da pessoa humana.

O ano de 1947 ficou marcado na história das ciências pela assinatura do Código de Nuremberg, instrumento do Direito Público Internacional que vinculava os seus signatários a observar, no âmbito dos seus territórios, uma série de procedimentos éticos no que tange ao desenvolvimento de pesquisas nas quais estivessem envolvidos seres humanos. A consolidação dos fundamentos contidos no Código de Nuremberg se deu, sem dúvida, com a assinatura da Declaração Universal dos Direitos Humanos, em 1948, que, por sua envergadura e adesão de vários estados, disseminou princípios importantes, insculpidos nas constituições das nações signatárias. A dignidade da pessoa humana, os princípios da igualdade e da liberdade e a inalienabilidade de direitos, tal qual a vida, foram, sem dúvida alguma, os que trouxeram maiores impactos para a normatização internacional das pesquisas científicas no período do pós-guerra.

Outra contribuição da Declaração Universal dos Direitos Humanos que trouxe grande repercussão entre as ciências humanas e sociais no século XX, e principalmente nestas primeiras décadas do século XXI, é o reconhecimento dos direitos das minorias étnicas, nacionais e religiosas, bem como a declaração de existência de direitos culturais e linguísticos inerentes a todas as comunidades humanas. O

reconhecimento desses direitos tem possibilitado extrair leituras que transcendam o aspecto biomédico contido nas intenções do Código de Nuremberg.

No Brasil, a porta de entrada das normas internacionais que visavam regular os procedimentos de pesquisa, após sua recepção constitucional, não se deu por intermédio do Poder Legislativo ou pelo Ministério de Ciência e Tecnologia, como poderíamos, inadvertidamente, supor. O fato de terem o seu nascedouro vinculado às pesquisas médicas envolvendo seres humanos fez com que ficasse a cargo do Ministério da Saúde a regulamentação da questão no país, materializada por meio do seu Conselho Nacional de Saúde, através da Resolução n. 196, de 10 de outubro de 1996.

Apesar de não ter sido gerada no seio do Poder Legislativo e não se constituir, portanto, em lei, a Resolução n. 196/96, já no seu preâmbulo, fundamentava-se nas principais declarações e diretrizes internacionais que regulam a matéria e cumpria as disposições dos principais diplomas normativos do país, tais quais a Constituição da República Federativa do Brasil, o Código Civil, o Código Penal, o Estatuto da Criança e do Adolescente, entre outros, o que lhe deu força de norma cogente[1].

Um avanço importante trazido pela Resolução n. 196/96 foi a criação da Comissão Nacional de Ética em Pesquisa (CONEP/MS), uma instância colegiada, de natureza consultiva, deliberativa, normativa, educativa, independente e vinculada ao Conselho Nacional de Saúde. Para além da CONEP/MS, foram também criados os Comitês de Ética em Pesquisa, órgãos colegiados que devem ser implementados, preferencialmente, nas instituições que realizam pesquisas envolvendo seres humanos.

Como estava previsto no seu texto preambular, a Resolução n. 196/96 deveria passar por revisões periódicas para que pudesse ser adequada às novas demandas, principalmente na área das tecnologias e das questões éticas. Fruto desse processo de revisão, o Plenário do Conselho Nacional de Saúde, em sua 240ª reunião ordinária, realizada nos dias 11 e 12 de dezembro de 2012, aprovou as novas normas regulamentadoras de pesquisas envolvendo seres humanos: a Resolução n. 466, de 12 de dezembro de 2012, publicada no Diário Oficial da União em 13 de junho de 2013 e que revoga, integralmente, as Resoluções CNS 196/96, 303/2000 e 404/2008.

Fiel às suas origens, o novo texto mantém-se ainda excessivamente biomédico, mas aponta, em seu item XIII.3, o fato de que as especificidades éticas das pesquisas nas ciências sociais e humanas e de outras que se utilizam de metodologias próprias dessas áreas serão contempladas em resolução complementar, dadas suas particularidades.

1 Norma cogente é aquela que constrange a quem se aplica, tornando seu cumprimento obrigatório de maneira coercitiva.

2. "ÉTICA IGUAL, PESQUISAS DIFERENTES"

Com esse título, em um texto da sua autoria, publicado no portal Ciência Hoje, o antropólogo e pesquisador Luiz Fernando Dias Duarte, da Universidade Federal do Rio de Janeiro, discutia, antes mesmo de entrar em vigor a Resolução n. 466/12, a situação das ciências sociais e humanas e apontava a grande dissonância com o texto da resolução mãe, a n. 196/96, no que tange às especificidades dessas ciências.

> Concebida precipuamente para lidar com as situações de pesquisa na área médica (e suas tecnologias), a resolução acabou se propondo a regular todas as pesquisas envolvendo 'seres humanos', mesmo aquelas cujas características nada têm de tecnológicas ou interventivas, como as da sociologia, da psicologia (não experimental) e da antropologia. Uma rede de Comitês de Ética em Pesquisa (CEP), subordinados a uma Comissão Nacional de Ética em Pesquisa (CONEP) vinculada ao Ministério da Saúde, foi criada em todo o país com atribuições universais de controle e fiscalização dos projetos de pesquisa. (DUARTE, 2009).

Ainda segundo Duarte (2009), essa seria uma temática de solução relativamente simples, caso fosse realizada a distinção entre os termos pesquisa *em* seres humanos e pesquisa *com* seres humanos. Enquanto as ciências médicas realizam, na maioria dos casos, pesquisas *em* seres humanos, de caráter interventivo, as demais ciências humanas e sociais realizam pesquisas *com* seres humanos, sem lhes afetar diretamente as condições de saúde, a incolumidade física ou quaisquer outros aspectos de natureza biológica.

Atualmente, a Associação Brasileira de Antropologia (ABA), com o aval da Associação Nacional de Pós-Graduação e Pesquisa em Ciências Sociais (ANPOCS) e de várias associações representantes das diversas áreas da pesquisa em ciências sociais e humanas, a exemplo da Associação Nacional de Pós-Graduação e Pesquisa em Letras e Linguística (ANPOLL), vem, através de um Grupo de Trabalho constituído também por membros da CONEP/MS, discutindo e delineando o documento que complementará a Resolução n. 466/12 nos aspectos que lhes forem pertinentes. Apesar dos avanços, as associações têm sido unânimes em alertar para o fato de que uma resolução complementar, vinculada à Resolução n. 466/12, como propõe essa norma, manterá as ciências humanas e sociais em uma situação de descabida subordinação às ciências médicas, tolhendo-lhes a autonomia metodológica. Desse modo, paralelamente ao fato de se elaborar um documento que seja capaz de normatizar as pesquisas nas humanidades, o Grupo de Trabalho vem tentando desvincular as suas normas éticas do âmbito do Ministério da Saúde para o Ministério da Ciência e Tecnologia, com a criação de comitês específicos para avaliar e autorizar os projetos de pesquisa oriundos dessas áreas do conhecimento.

3. ASPECTOS LEGAIS ENVOLVIDOS NA COLETA DE DADOS LINGUÍSTICOS

Em Linguística, muito embora não de forma exclusiva, os métodos de coleta de dados para a constituição dos *corpora* para as pesquisas frequentemente demandam o contato entre o pesquisador e indivíduos ou comunidades. Neste texto, focalizamos exclusivamente essa tipologia de coleta, relacionando-a às implicações legais à luz dos principais diplomas normativos do ordenamento jurídico brasileiro, bem como da própria Resolução n. 466/12 do Ministério da Saúde que, como já dissemos, apesar de não ser uma lei, tem valor de norma cogente.

Preliminarmente, conforme o preâmbulo da Resolução n. 466/12 apresenta, toda e qualquer pesquisa (entre elas a pesquisa linguística) deverá ter em mente os Princípios Fundamentais e os Direitos e Garantias Individuais elencados pela Constituição da República Federativa do Brasil, com destaque para a cidadania, a dignidade da pessoa humana, a igualdade e a vedação a toda e qualquer prática discriminatória que atente para esses direitos.

Uma grande diferença entre as Resoluções 196/96 e 466/12 diz respeito ao fato de que, na primeira, havia a declaração das principais leis brasileiras que deveriam ser observadas pelo pesquisador no curso da sua pesquisa envolvendo seres humanos; na última, apenas a Constituição é explicitamente citada e os demais diplomas foram substituídos pela expressão "Considerando a legislação brasileira correlata e pertinente".

O texto da Resolução n. 466/12, por sua vez, apresenta traços marcantes de várias leis brasileiras. Entretanto, nos interessam, de forma mais pontual, as menções ao Código Civil Brasileiro, ao Estatuto da Criança e do Adolescente, ao Estatuto do Idoso, bem como ao Código de Defesa do Consumidor.

O cuidado do linguista, em processo de coleta de dados, em relação ao Código Civil Brasileiro, deve focar principalmente no que concerne à produção de danos aos sujeitos participantes da pesquisa. Assim temos, conforme os artigos n. 186 e 187 da Lei n. 10.406, de 10 de janeiro de 2002:

> Art. 186 CC – Aquele que, por ação ou omissão voluntária, negligência ou imprudência, violar direito e causar dano a outrem, ainda que exclusivamente moral, comete ato ilícito.
> Art. 187 CC – Também comete ato ilícito o titular de um direito que, ao excedê-lo, excede manifestamente os limites impostos pelo seu fim econômico ou social, pela boa-fé ou pelos bons costumes.

Entre outras questões que possam ocorrer, a atenção ao cumprimento desses dois artigos do Código Civil na coleta de dados linguísticos remonta diretamente à aceitação explícita dos participantes da pesquisa, não se admitindo a

participação tácita. Há, ainda, a necessidade de observação do sigilo dos dados coletados e, principalmente, a proteção das identidades dos participantes. Esse aspecto é bem elucidado no próprio documento do Conselho Nacional de Saúde quando, no item V.7, prevê:

> Os participantes da pesquisa que vierem a sofrer qualquer tipo de dano resultante de sua participação na pesquisa, previsto ou não no Termo de Consentimento Livre e Esclarecido, têm direito à indenização, por parte do pesquisador, do patrocinador e das instituições envolvidas nas diferentes fases da pesquisa. (CONSELHO NACIONAL DE SAÚDE, 2012).

O Código de Defesa do Consumidor (CDC) traz uma contribuição importantíssima na proteção dos indivíduos e dos grupos que são submetidos às pesquisas de qualquer área do conhecimento, inclusive linguística, com a noção de vulnerabilidade. Pode até nos causar certo estranhamento a utilização do CDC às pesquisas que não se configuram relações de consumo; o fato é que o CDC apenas empresta a noção de vulnerabilidade para ser aplicada às relações entre as instituições, pesquisadores e os sujeitos pesquisados.

Essa noção de vulnerabilidade também está fortemente presente no Estatuto da Criança e do Adolescente e no Estatuto do Idoso, que colocam a criança, o adolescente e o idoso como naturalmente vulneráveis. No caso da Resolução n. 466/12, a vulnerabilidade é característica intrínseca dos sujeitos participantes da pesquisa.

> A eticidade da pesquisa implica em:
> – respeito ao participante da pesquisa em sua dignidade e autonomia, reconhecendo sua **vulnerabilidade**, assegurando sua vontade de contribuir e permanecer, ou não, na pesquisa, por intermédio de manifestação expressa, livre e esclarecida (CONSELHO NACIONAL DE SAÚDE, 2012, grifo nosso).

Para além desses diplomas normativos que já elucidamos, a própria Resolução n. 466/12 elenca um conjunto de fundamentos éticos e científicos pertinentes que devem ser obedecidos na condução de coleta de dados pelos pesquisadores. Vejamos:

> As pesquisas, em qualquer área do conhecimento envolvendo seres humanos, deverão observar as seguintes exigências:
> – ser adequada aos princípios científicos que a justifiquem e com possibilidades concretas de responder a incertezas;
> – estar fundamentada em fatos científicos, experimentação prévia e/ou pressupostos adequados à área específica da pesquisa;

– ser realizada somente quando o conhecimento que se pretende obter não possa ser obtido por outro meio;
– buscar sempre que prevaleçam os benefícios esperados sobre os riscos e/ou desconfortos previsíveis;
– utilizar os métodos adequados para responder às questões estudadas, especificando-os, seja a pesquisa qualitativa, quantitativa ou quali-quantitativa;
– se houver necessidade de distribuição aleatória dos participantes da pesquisa em grupos experimentais e de controle, assegurar que, a priori, não seja possível estabelecer as vantagens de um procedimento sobre outro, mediante revisão de literatura, métodos observacionais ou métodos que não envolvam seres humanos;
– obter consentimento livre e esclarecido do participante da pesquisa e/ou seu representante legal, inclusive nos casos das pesquisas que, por sua natureza, impliquem justificadamente, em consentimento *a posteriori***;**
– contar com os recursos humanos e materiais necessários que garantam o bem-estar do participante da pesquisa, devendo o(s) pesquisador(es) possuir(em) capacidade profissional adequada para desenvolver sua função no projeto proposto;
– prever procedimentos que assegurem a confidencialidade e a privacidade, a proteção da imagem e a não estigmatização dos participantes da pesquisa, garantindo a não utilização das informações em prejuízo das pessoas e/ou das comunidades, inclusive em termos de autoestima, de prestígio e/ou de aspectos econômico-financeiros;
– ser desenvolvida preferencialmente em indivíduos com autonomia plena. Indivíduos ou grupos vulneráveis não devem ser participantes de pesquisa quando a informação desejada possa ser obtida por meio de participantes com plena autonomia, a menos que a investigação possa trazer benefícios aos indivíduos ou grupos vulneráveis;
– respeitar sempre os valores culturais, sociais, morais, religiosos e éticos, como também os hábitos e costumes, quando as pesquisas envolverem comunidades;
– garantir que as pesquisas em comunidades, sempre que possível, traduzir-se-ão em benefícios cujos efeitos continuem a se fazer sentir após sua conclusão. Quando, no interesse da comunidade, houver benefício real em incentivar ou estimular mudanças de costumes ou comportamentos, o protocolo de pesquisa deve incluir, sempre que possível, disposições para comunicar tal benefício às pessoas e/ou comunidades;
– comunicar às autoridades competentes, bem como aos órgãos legitimados pelo Controle Social, os resultados e/ou achados da pesquisa, sempre que estes puderem contribuir para a melhoria das condições de vida da coletividade, preservando, porém, a imagem e assegurando que os participantes da pesquisa não sejam estigmatizados;
– assegurar aos participantes da pesquisa os benefícios resultantes do projeto, seja em termos de retorno social, acesso aos procedimentos, produtos ou agentes da pesquisa;
– utilizar o material e os dados obtidos na pesquisa exclusivamente para a finalidade prevista no seu protocolo, ou conforme o consentimento do participante; (CONSELHO NACIONAL DE SAÚDE, 2012, grifo nosso).

4. DO PROCESSO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Ponto fulcral do texto da Resolução n. 466/12 diz respeito ao Processo de Consentimento Livre e Esclarecido. O respeito ao princípio da dignidade da pessoa humana exige que toda pesquisa se processe com consentimento livre e esclarecido dos participantes, indivíduos ou grupos que, por si e/ou por seus representantes legais, manifestem a sua anuência à participação no procedimento.

Entende-se por Processo de Consentimento Livre e Esclarecido o procedimento através do qual o sujeito convidado a participar de uma pesquisa seja devidamente elucidado acerca das etapas a serem percorridas, para que possa se manifestar, de forma autônoma, consciente, livre e esclarecida. Essa etapa da coleta de dados sofreu modificações em relação ao texto normativo da Resolução n. 196/96 por conta de pressões advindas principalmente dos pesquisadores das ciências humanas e sociais que – enfrentando aquilo que Labov (2008) chamou, na Sociolinguística, de *paradoxo do observador* – tinham as suas pesquisas prejudicadas pelo fato de contaminar a coleta de dados por conta da necessidade de obtenção prévia do Termo de Consentimento Livre e Esclarecido.

> A etapa inicial do Processo de Consentimento Livre e Esclarecido é a do esclarecimento ao convidado a participar da pesquisa, ocasião em que o pesquisador, ou pessoa por ele delegada e sob sua responsabilidade, deverá:
> – buscar o momento, condição e local mais adequados para que o esclarecimento seja efetuado, considerando, para isso, as peculiaridades do convidado a participar da pesquisa e sua privacidade;
> – prestar informações em linguagem clara e acessível, utilizando-se das estratégias mais apropriadas à cultura, faixa etária, condição socioeconômica e autonomia dos convidados a participar da pesquisa; e
> – conceder o tempo adequado para que o convidado a participar da pesquisa possa refletir, consultando, se necessário, seus familiares ou outras pessoas que possam ajudá-los na tomada de decisão livre e esclarecida. (Conselho Nacional de Saúde, 2012).

Superada a etapa inicial de esclarecimento, o pesquisador responsável, ou pessoa por ele delegada, deverá apresentar ao convidado para participar da pesquisa, ou a seu representante legal, o Termo de Consentimento Livre e Esclarecido para que seja lido e compreendido, antes da concessão do seu consentimento livre e esclarecido. Nele, obrigatoriamente deverão estar contidos:

> – justificativa, os objetivos e os procedimentos que serão utilizados na pesquisa, com o detalhamento dos métodos a serem utilizados, informando a possibilidade de inclusão em grupo controle ou experimental, quando aplicável;

– explicitação dos possíveis desconfortos e riscos decorrentes da participação na pesquisa, além dos benefícios esperados dessa participação e apresentação das providências e cautelas a serem empregadas para evitar e/ou reduzir efeitos e condições adversas que possam causar dano, considerando características e contexto do participante da pesquisa;
– esclarecimento sobre a forma de acompanhamento e assistência a que terão direito os participantes da pesquisa, inclusive considerando benefícios e acompanhamentos posteriores ao encerramento e/ ou a interrupção da pesquisa;
– garantia de plena liberdade ao participante da pesquisa, de recusar-se a participar ou retirar seu consentimento, em qualquer fase da pesquisa, sem penalização alguma;
– garantia de manutenção do sigilo e da privacidade dos participantes da pesquisa durante todas as fases da pesquisa;
– garantia de que o participante da pesquisa receberá uma via do Termo de Consentimento Livre e Esclarecido;
– explicitação da garantia de ressarcimento e como serão cobertas as despesas tidas pelos participantes da pesquisa e dela decorrentes; e
– explicitação da garantia de indenização diante de eventuais danos decorrentes da pesquisa. (CONSELHO NACIONAL DE SAÚDE, 2012)

O Termo de Consentimento Livre e Esclarecido deverá, ainda:

– conter declaração do pesquisador responsável que expresse o cumprimento das exigências contidas no item IV. 3 da Resolução n. 466/12;
– ser aprovado pelo CEP perante o qual o projeto foi apresentado e pela CONEP, quando pertinente; e
– ser elaborado em duas vias, rubricadas em todas as suas páginas e assinadas, ao seu término, pelo convidado a participar da pesquisa, ou por seu representante legal, assim como pelo pesquisador responsável, ou pela (s) pessoa (s) por ele delegada (s), devendo as páginas de assinaturas estar na mesma folha. Em ambas as vias deverão constar o endereço e contato telefônico ou outro, dos responsáveis pela pesquisa e do CEP local e da CONEP, quando pertinente. (CONSELHO NACIONAL DE SAÚDE, 2012).

Por fim, a Resolução n. 466/12 trata dos indivíduos em restrição da liberdade ou do esclarecimento necessário para o adequado consentimento: em pesquisas nas quais a coleta de dados é realizada com crianças e adolescentes, estudantes, militares, presidiários e pessoas relativamente ou absolutamente incapazes, na acepção do Código Civil, ou ainda nos casos em que os agrupamentos humanos estejam submetidos à líderes religiosos ou de outra natureza, como também, no caso dos índios, cuja submissão ao ato autorizativo prévio é para a FUNAI.

– em pesquisas cujos convidados sejam crianças, adolescentes, pessoas com transtorno ou doença mental ou em situação de substancial diminuição em sua capacidade de decisão, deverá haver justificativa clara de sua escolha, especificada no protocolo e aprovada pelo

> CEP, e pela CONEP, quando pertinente. Nestes casos deverão ser cumpridas as etapas do esclarecimento e do consentimento livre e esclarecido, por meio dos representantes legais dos convidados a participar da pesquisa, preservado o direito de informação destes, no limite de sua capacidade;
> – a liberdade do consentimento deverá ser particularmente garantida para aqueles participantes de pesquisa que, embora plenamente capazes, estejam expostos a condicionamentos específicos, ou à influência de autoridade, caracterizando situações passíveis de limitação da autonomia, como estudantes, militares, empregados, presidiários e internos em centros de readaptação, em casas-abrigo, asilos, associações religiosas e semelhantes, assegurando-lhes inteira liberdade de participar, ou não, da pesquisa, sem quaisquer represálias;
> – em comunidades cuja cultura grupal reconheça a autoridade do líder ou do coletivo sobre o indivíduo, a obtenção da autorização para a pesquisa deve respeitar tal particularidade, sem prejuízo do consentimento individual, quando possível e desejável.
>
> Quando a legislação brasileira dispuser sobre competência de órgãos governamentais, a exemplo da Fundação Nacional do Índio – FUNAI, no caso de comunidades indígenas, na tutela de tais comunidades, tais instâncias devem autorizar a pesquisa antecipadamente. (CONSELHO NACIONAL DE SAÚDE, 2012).

Nos casos em que é inviável a obtenção do Termo de Consentimento Livre e Esclarecido ou que sua obtenção signifique riscos substanciais à privacidade e confidencialidade dos dados do participante ou aos vínculos de confiança entre pesquisador e pesquisado, a sua dispensa deve ser justificadamente solicitada pelo pesquisador responsável ao Sistema CEP/CONEP, para apreciação, sem prejuízo do posterior processo de esclarecimento.

RECOMENDAÇÕES FINAIS

A coleta de dados para constituição de *corpora* para pesquisas em linguística, bem como nas demais ciências humanas e sociais, está subordinada às questões ético-legais constantes em diplomas normativos brasileiros, além de estar diretamente vinculada às normas emanadas pelo Conselho Nacional de Saúde, através da Resolução n.466 de 12 de dezembro de 2012.

O direcionamento biomédico que historicamente consta nas resoluções vem despertando movimentos, no interior das humanidades, no sentido de que sejam criadas normas específicas e que, para além disso, ocorra a criação de um Conselho de Ética para analisar as pesquisas das ciências humanas e sociais, no seio do Ministério da Ciência e Tecnologia, e não do Ministério da Saúde, conforme o modelo atualmente vigente.

Apesar de constar na Resolução n. 466/12 um dispositivo que autoriza a elaboração de uma resolução complementar que atenda às especificidades das ciências sociais e humanas, não significa que essas ciências não devam obedecer aos preceitos constantes do texto principal. A atenção aos princípios constitucionais da dignidade da pessoa humana, da cidadania, dos valores sociais e da vedação ao preconceito devem sempre nortear a coleta de dados que envolvam diretamente seres humanos.

REFERÊNCIAS

CONSELHO NACIONAL DE SAÚDE. Resolução n. 196, de 10 de outubro de 1996. Dispõe sobre diretrizes e normas regulamentadoras de pesquisas envolvendo seres humanos. Disponível em: <http://conselho.saude.gov.br/web_comissoes/conep/aquivos/resolucoes/resolucoes.htm>. Acesso em: 02 fev. 2014. CONSELHO NACIONAL DE SAÚDE. Resolução n. 466, de 12 de dezembro de 2012. Dispõe sobre diretrizes e normas regulamentadoras de pesquisas envolvendo seres humanos. Disponível em: <http://conselho.saude.gov.br/resolucoes/2012/Reso466.pdf>. Acesso em: 10 fev. 2014.

BRASIL. *Constituição* (1988). Contêm as emendas constitucionais posteriores. Brasília, DF: Senado, 2013.

DUARTE, L. F. D. Ética igual pesquisas diferentes. *Instituto Ciência Hoje*, set 2009. Disponível em: <http://cienciahoje.uol.com.br/colunas/sentidos-do-mundo/etica-igual-pesquisas-diferentes>. Acesso em: fev. 2014.

CAPÍTULO 2

ASPECTOS TÉCNICOS NA COLETA DE DADOS LINGUÍSTICOS ORAIS

Miguel Oliveira, Jr.

INTRODUÇÃO

Existem conjuntos de práticas específicas para a coleta de dados orais que vêm sendo adotados internacionalmente em projetos de documentação linguística. Órgãos como E-MELD School of Best Practice, Open Language Archives Community (OLAC) e o Comitê Técnico da International Association of Sound and Audiovisual Archives (IASA), estudam e propõem tais práticas. É, todavia, ainda um fato recorrente a não observação dessas recomendações técnicas na coleta de dados para estudos linguísticos, seja pela falsa ideia de que o mais importante é o dado de fala em si e não a qualidade do registro, ou simplesmente pela falta de conhecimento técnico do assunto (VAUX; COOPER, 1999).

O objetivo do presente texto é apresentar de maneira concisa, usando uma linguagem acessível, técnicas recentes de recolha de dados orais para documentação linguística e listar recomendações básicas para a sua adequada realização. Não se intenta aqui explorar métodos teóricos de recolha propriamente, algo que vai depender, evidentemente, dos objetivos de pesquisa individuais, mas oferecer sugestões técnicas, baseadas em recomendações feitas por órgãos internacionais de codificação e transmissão de dados de áudio.

1. RECOLHA DE DADOS ORAIS: CONSIDERAÇÕES INICIAIS

A escolha de equipamentos de gravação de áudio para a construção de um *corpus* de fala deve ser guiada por estratégias que garantam um bom resultado do material a ser coletado. A coleta de dados é a base de um bom *corpus*.

A seguir, veremos alguns dos princípios essenciais de gravação de dados de fala que visam a melhoria de sua precisão, de sua consistência e de sua confiabilidade. O que se deve sempre considerar na coleta de dados é que o material coletado pode ser útil para diversos estudos futuros, inclusive estudos que exijam uma alta qualidade dos dados (como, por exemplo, determinadas análises acústicas). É importante, por esse motivo, sempre primar pela qualidade na coleta de dados.

2. GRAVAÇÃO DE ÁUDIO

É fundamental, por razões já expostas, garantir que o arquivo de áudio gravado contenha o maior número possível de informações do sinal original. Essa regra vale tanto para gravações quanto para digitalizações de uma gravação analógica existente.

Em princípio, um gravador deve ser capaz de capturar a resposta de frequência (<10.000Hz) e o alcance dinâmico (40dB), característicos da fala. No entanto, existem outras considerações importantes ao escolher um gravador para coleta de dados de fala.

2.1. Gravação Analógica

A gravação analógica pode capturar a resposta de frequência e o alcance dinâmico, próprios da fala, com precisão e detalhe, sobretudo se forem utilizados um bom microfone pré-amplificado e um gravador de fita magnética profissional. No entanto, os gravadores analógicos geralmente vêm acompanhados de mecanismos de transporte ruidosos, não oferecem a possibilidade de registrar o tempo, o que dificulta a análise e o manuseio dos dados, utilizam um meio (a fita analógica) bastante frágil e, o mais importante: para fazer parte de arquivos digitais (algo altamente recomendável, como veremos mais adiante), as gravações precisam ser digitalizadas, o que implica em perda de informações. É importante ressaltar que equipamentos analógicos e mídias analógicas são cada vez mais difíceis de serem encontrados no mercado, o que torna difícil esse tipo de gravação.

2.2. Gravação digital

Bastante comuns e acessíveis hoje em dia, os gravadores digitais podem ser encontrados nos mais diversos modelos, com características também diversificadas: DAT, *Minidisk*, *solid state*, CD/DVD-R, e *hard disk* são alguns exemplos. O que basicamente os distingue são o meio e o formato de gravação. Algumas dessas tecnologias, como o DAT e o *Minidisk*, são hoje consideradas obsoletas. A tendência atual é a utilização de gravadores *solid state*. Esses gravadores utilizam como mídia cartões de memória estável (*secure digital*), o que os torna menos vulneráveis a choques e vibrações, e desse modo, a interferências/ruídos no sinal registrado.

Na construção de um *corpus* de fala, a recomendação é que os gravadores registrem o áudio em formato não comprimido. Na hora de decidir pelo equipamento adequado, é importante considerar as taxas de amostragem e resolução com que trabalha (idealmente 24 bit/96kHz), as entradas para microfone (XLR), o sistema de alimentação (alimentação fantasma, pilhas, USB), a resposta em frequência de (20Hz - 20kHz) e o alcance dinâmico (> 80dB). É muito importante que o gravador faça registro em formato não comprimido (WAV, por exemplo).

Os gravadores portáteis Marantz da linha PMD, como o PMD6621, estão entre os mais utilizados em coleta de dados orais que exijam alguma qualidade. São modelos portáteis que fazem gravação em mídia SD e têm entrada XLR para microfones profissionais. Pela metade do preço do Marantz PMD661, o gravador Sony PCM-M10 também é uma boa opção para projetos menos exigentes. A utilização de um bom microfone (como o Audix HT5, por exemplo) poderá, nesse caso, fazer a diferença. Como só há entrada de microfone de 3,5 mm, será preciso adquirir um adaptador XLR-para-3,5mm para usar microfones profissionais.

3. MICROFONES

O microfone é parte fundamental na recolha de dados orais, senão o mais importante. Escolher o microfone que melhor se adeque aos propósitos da recolha e às especificações do equipamento de gravação exige algum conhecimento básico acerca da estrutura e mecanismo de operação desses equipamentos. Os microfones são descritos pelo seu princípio transdutor, por sua direcionalidade e pelo tipo de uso a que se destinam (*design*). Assim, os microfones podem ser dinâmicos, condensadores (capacitivos), omnidirecionais, bidirecionais, cardioides, hipercardioides, *shotguns*, de lapela, de mão, *headsets*, etc.

Os microfones dinâmicos são bastante resistentes, não precisam de baterias ou fontes de alimentação externas (o que é uma vantagem), mas não vêm equipados

com pré-amplificadores. Por conta de suas características, podem não responder bem aos sons agudos e/ou transientes, próprios da voz, o que os torna pouco indicados para gravações da fala.

Os microfones condensadores respondem com clareza a sons transientes, o que resulta em sons mais naturais, limpos e ricos em detalhes. Além disso, pesam muito menos e podem ser muito menores. No entanto, porque requerem uma fonte de energia e têm cápsulas muito mais delicadas, podem apresentar algumas dificuldades de uso e maior cuidado de manipulação.

Os microfones também podem ser identificados por suas propriedades direcionais, isto é, pela forma como foram planejados para captar o som de diversas direções. Assim, os microfones podem ser classificados geralmente como pertencendo a um de dois grupos principais: omnidirecional e direcional. Os microfones omnidirecionais captam igualmente sons de todas as direções. Os direcionais podem ter várias configurações, sendo as mais comuns os cardioides, que captam sons provenientes da frente, e os bidirecionais, que registram sons que vêm de direções opostas. Um grande problema de microfones direcionais, como os cardioides, é o chamado efeito de proximidade. O efeito de proximidade refere-se ao aumento na resposta de frequências baixas em função da proximidade da fonte sonora, o que pode causar problemas em análises acústicas. Para evitar o efeito de proximidade, recomenda-se o uso de microfones omnidirecionais.

Em se tratando do *design*, talvez o tipo de microfone mais popular é o chamado microfone de mão: são mais naturais e dão maior liberdade aos usuários; são geralmente utilizados em situações que exigem a troca de falantes, como em entrevistas. Entretanto, são muito sensíveis à ruídos de manuseio e podem ser cansativos se usados por um longo período de tempo. Os microfones de lapela são usados em situações que requerem de seus usuários as mãos desocupadas. Oferecem maior liberdade e espontaneidade, mas são também muito suscetíveis a ruídos provocados por roupa, cabelo e batidas acidentais no microfone ou no cabo. Os microfones *headset* são utilizados quando se quer garantir uma distância fixa e consistente entre o microfone e a boca do usuário, oferecendo ao mesmo tempo uma relativa liberdade de locomoção. Têm sido considerados a melhor alternativa para documentação linguística (PLICHTA, 2004).

Antes de selecionar um microfone, é preciso também levar em conta as suas características técnicas. Além do princípio transdutor, da direcionalidade e do design é preciso atenção para: (i) a impedância (que deve ser menor que 600Ω); (ii) a resposta em frequência (que deve ser entre 20Hz – 20kHz, plana); (iii) o nível máximo de pressão sonora (> 120dB); (iv) a sensibilidade (média, por volta de 6mV/Pa); (v) o ruído próprio (< 32dBA) e (vi) o tipo de conexão. Essas informações são disponibilizadas pelo fabricante, mas podem ser obtidas em organizações independentes, como a Microphone Data <http://www.microphone-data.com/>.

O DPA 4006-TL é um dos melhores microfones de mão do mercado para gravação de voz devido à sua alta sensibilidade, excelente resposta em frequência (sobretudo as frequências baixas) e baixíssimo ruído próprio. Por conta de suas características, deve ser usado em ambientes silenciosos, com suporte apropriado e, de preferência, com proteção de explosões (*pop filter*). Se a gravação exige microfone de lapela, o Sanken COS-11DBP é uma excelente opção: o microfone pode ser usado com bateria ou com alimentação fantasma, registra uma vasta gama de frequências e é bastante discreto. O uso de microfones de lapela, no entanto, deve ser considerado com bastante cautela se o objetivo for registrar áudio para análise linguística. Como já apontado acima, esse tipo de microfone capta ruídos indesejados e não registra informações acústicas importantes, por conta de sua localização em relação à boca do entrevistado (discutiremos esta questão em particular logo mais abaixo).

Considerado um dos melhores *headsets* no mercado para registro de voz, o DPA 4066 tem uma resposta em frequência ampla e plana, padrão ominidirecional (o que permite o registro em proximidades muito curtas), média sensibilidade e habilidade de lidar com altos níveis de pressão, o que possibilita a captura de sinais ricos em detalhes, com amplo *headroom*. Conecta-se a praticamente qualquer tipo de entrada, com o adaptador adequado. O Audix HT5, por sua vez, é um *headset* relativamente barato (sobretudo se comparado ao DPA 4066). Possui um *design* bastante confortável e apropriado, um adaptador que provê energia independente (com baterias do tipo AA) e média sensibilidade, o que é ideal para gravações de fala natural, evitando o uso de altos ganhos do pré-amplicador/alimentação fantasma e, assim, possíveis ruídos indesejáveis. A relação s/n é bastante boa, assim como a resposta em baixas frequências. Bastante usado em pesquisa acústica, o Opus 55 Mk II tem sido muito elogiado pela crítica especializada por sua habilidade de apresentar respostas muito planas (resultando em um envelope de fala bastante realista) e amplas (registrando harmônicos perto da fundamental). Ao contrário da maioria dos *headsets*, que têm baixa sensibilidade para registrar altos volumes, o Opus 55 tem sensibilidade média, o que é apropriado para voz em altura normal. Requer um adaptador para usar com alimentação fantasma.

4. FORMATO DO ARQUIVO DE ÁUDIO

Na avaliação de equipamentos digitais a serem adquiridos, é preciso sempre levar em consideração o formato com que trabalham. Os equipamentos devem registrar as sessões de coleta de dados nos formatos recomendados por arquivos digitais com reputação internacional.

Recomenda-se que a escolha do formato digital baseie-se nos seguintes parâmetros (SIMONS, 2006):

- sem perdas: o formato deve garantir o registro integral do conteúdo;
- padrões abertos: o formato deve usar um padrão computacional que seja aberto;
- transparente: o formato deve ser acessível;
- usado em muitos aplicativos: é preciso garantir que o arquivo seja aberto no futuro.

O formato não comprimido WAV tem se tornado padrão em arquivos destinados a pesquisas linguísticas, sendo o mais recomendado por órgãos de referência e, em muitos casos, o único formato aceito para preservação em bancos de dados internacionais.

5. RECOMENDAÇÕES PARA GRAVAÇÃO DE ÁUDIO

No que diz respeito à gravação de dados propriamente, um dos primeiros cuidados que se deve ter é estabelecer níveis corretos de registro de áudio. Trata-se de um procedimento que deve ser feito pontualmente, ou seja, não é possível estabelecer um nível padrão para todas as situações de gravação.

A regra geral para definir os níveis de gravação é regular o ganho do amplificador em um nível que permita uma gravação em uma altura adequada, com um amplo *headroom*, sem causar sobrecarga e/ou clipagem. Essas recomendações servem para gravações em que o microfone possa estar sempre perto da boca, ou com microfone pré-amplificados de alta qualidade.

Em gravações feitas em salas silenciosas ou acusticamente tratadas, recomenda-se gravar o mais baixo possível, deixando um amplo *headroom* que possa ser manipulado posteriormente, caso se faça necessário. Em ambientes ruidosos, com um microfone montado à cabeça, recomenda-se a gravação em um nível de -12dBFS para garantir uma relação sinal-ruído favorável e detalhes espectrais adequados. Se o microfone pré-amplificado gerar muito ruído nesse nível, recomenda-se a diminuição do ganho em 6dB.

O processo de registro de dados orais deve estar amparado não apenas no conhecimento técnico dos equipamentos utilizados no registro em si, mas também – e sobretudo – no conhecimento da metodologia empregada no registro de voz. Para isso, é fundamental seguir as orientações sugeridas por órgãos reconhecidos, como, por exemplo, o Open Archival Information System (OAIS), que é um modelo de referência, com padrão ISO 14721:2003, adotado pelos bancos de

dados linguísticos mais recentes, e o Comitê Técnico da International Association of Sound and Audiovisual Archives (IASA).

A recomendação mínima que se dá para gravação digital de áudio é a utilização de uma taxa de amostragem igual ou superior a 44.100Hz, com uma resolução superior a 16bit, em formato não comprimido (.wav). Estudos recentes em psicoacústica, no entanto, revelaram que níveis de amostragem mais altos têm efeitos perceptuais significativos, o que aponta para a recomendação de uma taxa de amostragem mínima de 96kHz (PLICHTA, 2004). As gravações devem ser feitas em apenas um canal (monoaural).

5.1. Ruído

No momento da gravação, é preciso muita atenção para as seguintes possíveis fontes de ruído: (I) nas proximidades: pessoas, animais, atividades (domésticas ou outras); (II) nas redondezas: o tráfego, geradores, aviões; (III) máquinas: geladeiras, ventiladores, computadores, telefones celulares; (IV) não audível: interferência elétrica; (V) eventos imprevisíveis: farfalhar de papéis ou de roupa, mesa ou cadeira em desequilíbrio; (VI) eventos previsíveis (que podem ser evitados): tamborilar em mesa, *feedback* do pesquisador; (VII) manuseio do equipamento: mouse, teclados, microfone, cabos; (VIII) gerados por equipamentos: cabos inapropriados ou baratos, nível de sinal de entrada inapropriado, etc.

Há situações em que o ruído é inevitável. Nelas, é preciso encontrar uma solução para minimizar a sua interferência na gravação. Para evitar o trânsito, o ideal é investigar horários de pico e gravar apenas fora deles. Se não for possível, utilizar um microfone unidirecional. Para evitar interferência de fenômenos naturais (vento, chuva, trovão), use um "gato morto": acessório protetor de vento acoplado ao microfone. O "gato morto" pode também ser útil para lugares com motores, ou geradores intermitentes. Nesses casos, no entanto, o indicado é garantir que serão desligados ou procurar um outro lugar. Crianças e animais precisam ser monitorados. O ideal é investigar, quando possível, o melhor momento.

5.2. Distância do microfone

Um dos principais fatores para garantir um sinal de áudio de boa qualidade é a observação de uma distância mínima (e constante) entre a boca do participante e o microfone. A recomendação de distância mínima (e máxima) varia de acordo com as características e especificidades do microfone utilizado, mas, em geral, gira em torno de 30 cm a 15 cm para a maior parte dos microfones.

De acordo com a lei do inverso do quadrado, dobrando-se a distância entre o microfone e a fonte, a energia (ou intensidade) sonora registrada cai quatro vezes, aumentando as chances de se captar sinais indesejados. O ideal, portanto, é que o registro de voz deva ser feito com o microfone posicionado o mais próximo possível da boca do participante. Mas é preciso atenção para o efeito de proximidade, tal como observado acima.

Microfones do tipo *headset* ficam a uma distância relativamente curta da boca do participante (1-5cm), com a cápsula posicionada ao lado da boca (e não à frente ou abaixo). Para evitar o efeito de proximidade, a maioria dos microfones *headset* são omnidirecionais. Além disso, os melhores microfones desta categoria são equipados com *shock-mounts*, o que garante gravações livres de ruídos causados por choques mecânicos.

RECOMENDAÇÕES FINAIS

O presente texto teve por objetivo apresentar métodos recomendados para registro de dados orais em pesquisas linguística e discutir propriedades técnicas úteis para a escolha de equipamentos adequados, garantindo assim a qualidade dos dados coletados. Essas recomendações, todavia, não devem ser consideradas empecilhos para a recolha de dados em situações extremas. Como bem observa Bowern (2008), uma gravação feita em uma situação distante do ideal é melhor que nenhuma gravação. É preciso considerar que as condições de coleta de dados nunca serão perfeitas. Portanto, o que se recomenda é sempre planejar cuidadosamente a recolha de dados, e na hora, lançar mão do melhor que se puder.

REFERÊNCIAS

BOWERN, C. *Linguistic fieldwork*: a practical guide. Basingstoke: Palgrave Macmillan, 2008.

PLICHTA, B. Data aquisition problems. *Signal acquisition and acoustic analysis of speech*, 2004. Disponível em: <http://bartus.org/akustyk/signal_aquisition.pdf>. Acesso em: 11 ago. 2013.

SIMONS, G. F. Ensuring that digital data last: the priority of archival form over working form and presentation form. SIL *Electronic Working Papers 2006-003*. Boston: mar 2006.

VAUX, B.; COOPER, J. *Introduction to Linguistic Field Methods*. Munich: Lincom Europa, 1999.

CAPÍTULO 3

O PROJETO *A LÍNGUA PORTUGUESA NO SEMIÁRIDO BAIANO – FASE 3*: CRITÉRIOS DE CONSTITUIÇÃO E DA AMOSTRAGEM DO BANCO DE DADOS

Silvana Silva de Farias Araujo
Norma Lucia Fernandes de Almeida

INTRODUÇÃO

O projeto de pesquisa *A Língua Portuguesa no Semiárido Baiano* foi implementado no ano de 1996 (embora só tenha sido oficializado em 1998), na Universidade Estadual de Feira de Santana (UEFS), sob a coordenação das professoras Norma Almeida e Zenaide Carneiro. Inicialmente, as atividades do projeto centraram-se na formação de *corpora* gravados em comunidades rurais da região semiárida baiana. A gravação desse material começou em 1996 e estendeu-se até o ano de 2001, tendo sido contempladas comunidades localizadas em diferentes regiões sertanejas da Bahia. O critério utilizado para a realização da coleta de dados foi o de que as localidades apresentassem formações sócio-histórico e demográficas diferenciadas, fazendo parte, principalmente, de dois importantes fatores do processo de urbanização do interior do estado (ALMEIDA; CARNEIRO, 1999), a saber, os chamados ciclos da *agropecuária*, cujas origens remontam aos séculos XVII a XIX

– Jeremoabo e Feira de Santana –, e da *mineração*, com origens mais densamente vinculadas ao século XVIII – Rio de Contas e Caém. Esse *corpus* foi publicado, com o apoio da Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB), por Almeida e Carneiro (2008), e diversas análises linguísticas vêm sendo realizadas, mesmo antes da publicação, já tendo servido como base empírica para artigos, monografias, dissertações e tese de doutorado (ALMEIDA; CARNEIRO, 2014).

No ano de 2007, o projeto entrou numa nova fase, denominada *Fase 3*, quando as atenções voltaram-se para a zona urbana de Feira de Santana/BA. Após já ter se delineado um quadro do português falado em comunidades rurais baianas, cabia aos pesquisadores do projeto a desafiadora e necessária tarefa de traçar uma descrição sociolinguística do português falado em uma cidade tão múltipla como Feira de Santana. Iniciaram-se, assim, naquele ano, novamente com o apoio da UEFS e da FAPESB, as gravações na sede do município, que é o segundo do estado da Bahia em termos populacionais, ficando atrás apenas da capital. Para essa nova etapa, o projeto contou com a coordenação de mais duas professoras, Eliana Pitombo e Silvana Araujo, além da colaboração de voluntários e de dez bolsistas de Iniciação Científica. É sobre essa fase do projeto que se discorre neste texto.

1. FEIRA DE SANTANA: CARACTERÍSTICAS SÓCIO-HISTÓRICO E DEMOGRÁFICAS

Na denominação do município, à 108 km de Salvador, subjaz muito das suas principais características. A palavra "Feira" remete à questão da diversidade, do conglomerado, do movimento, do colorido, do som, da circulação. Foi uma feira livre que proporcionou à Feira de Santana, ou simplesmente à Feira (como é comumente chamada), a ser o que é.

Talvez por sua localização singular, de fácil acesso, encontrando-se num dos principais entroncamentos de rodovias do norte-nordeste brasileiro, Feira de Santana reuniu conjunturas para vir a tornar-se a complexa cidade que é, com um "caldeirão demográfico" e com a presença de tantos contatos dialetais, como ocorre em poucas cidades interioranas do Brasil. Na sede do município, no seu perímetro urbano, passam as rodovidas BR 116 (Norte e Sul) e BR 324, enquanto no distrito de Humildes[1] passa a BR 101. A Figura 1 mostra a localização de municípios circunvizinhos e de

1 O município de Feira de Santana está dividido em bairros (na sede) e em distritos, esses últimos em número de oito: Bonfim de Feira, Governador João Durval Carneiro (antigamente, denominado Ipuaçu), Humildes, Jaquara, Jaíba, Maria Quitéria (antigamente, São José das Itapororocas), Matinha e Tiquaruçu. Matinha, antigo povoado do distrito de São José das Itapororocas, antigamente denominado Matinha dos Pretos, passou a ser considerado distrito de Feira de Santana a partir de 2008, com o Decreto n. 7.462, de 21 de fevereiro de 2008.

distritos, além do sítio da sede do município, circunscrita pelo Anel de Contorno. Devido à expansão imobiliária, acelerada nas últimas cinco décadas, há muitos bairros para além do Anel de Contorno de Feira de Santana.

Figura 1 – Representação espacial do Município de Feira de Santana. Fonte: <http://maps.google.com>. Acesso em: 24 out. 2011.

Diante dessas características espaciais, é muito comum pessoas de todas as regiões do Brasil terem ouvido falar ou já terem passado pelo município. No âmbito do estado da Bahia, os municípios circunvizinhos mantêm estreitas relações com Feira de Santana, estando os seus moradores em frequentes contatos entre si, alguns trabalhando e morando em cidades vizinhas, recorrendo à Feira de Santana quando precisam de serviços médicos, educacionais, comerciais e de lazer especializados.

Geograficamente, Feira de Santana localiza-se numa zona de transição entre o Recôncavo e o Semiárido, precisamente no agreste baiano[2], embora seja conhecida por "Princesa do Sertão", alcunha conferida por Ruy Barbosa quando de sua visita à cidade no ano de 1919. Apesar de não ser o foco deste texto, é preciso destacar que essa posição geográfica também reveste de especial interesse para o estudo do município em seus aspectos linguísticos. A Figura 2 aponta a posição intermediária de Feira de Santana (localizada na chamada Região do Paraguaçu), a meio caminho entre o Recôncavo – "o litoral acessível mais

2 Segundo Santos e Pinho (2003, p. 73), tradicionalmente "agreste" significa uma zona de transição entre a faixa litorânea e a zona semiárida. Atualmente, os estudiosos do assunto não fazem distinção conceitual entre *agreste* e *semiárido*, adotando unicamente a denominação semiárido.

próximo", na expressão de Neves (2008) – e o Sertão, funcionando como um portal para a região sertão/semiárida, algo que pode significar uma riqueza em suas normas linguísticas, vindo a abrigar características peculiares dos falares do interior e do litoral.

Figura 2 – Quinze regiões econômicas da Bahia. Fonte: SEI, 2003.

Vale destacar que Silva Neto (1963), ao tratar do período de formação do português brasileiro, traçou uma distinção ainda válida entre a língua da costa e a do interior (ARAUJO, S.; ARAUJO, J., 2009; ARAUJO, 2014), considerando-se a situação bipolarizada do seu contexto de formação, sobre a qual se explanará ainda neste texto.

A respeito da situação sócio-histórica do município de Feira de Santana, pode-se presumir uma situação de contatos linguísticos e culturais diversos. Houve uma intensa atividade pecuária e comercial que propiciou um grande tráfego de pessoas pelo sítio geográfico da cidade, destacando-se a figura do vaqueiro, certamente de origem indígena ou africana (negros libertos integrados nas relações

socioeconômicas), ou ainda, portuguesa de origem não nobre.[3] A propósito, destaca-se que, já nas suas origens, no final do século XVII[4], Feira de Santana caracterizava-se por ser um lugar de passagem de viajantes, vaqueiros e tropeiros, pois no seu território atual estava a Estrada das boiadas, por onde eram conduzidos animais comercializados em Cachoeira, Santo Amaro e Salvador.

Tal contexto sócio-histórico, demográfico e econômico sugere uma realidade polarizada no período de formação da variedade linguística feirense, tal como foi esboçada por Silva Neto (1963) e sistematizada por Lucchesi (1994; 2001; et al.) no que tange ao processo de formação do português brasileiro. Essa bipolarização pode ser associada às diferentes culturas em contato, destacando-se as das línguas dos indígenas, as dos escravos africanos e a do colonizador branco. No município, de um lado, havia fazendeiros, comerciantes, representantes da Igreja e do Estado e militares graduados, subordinados aos modelos advindos de Portugal; do outro, vaqueiros, roceiros, meeiros e escravos, que adquiriram o português como língua materna a partir de um modelo já adquirido como segunda língua por seus pais e livre de normatizações.

Embora, como bem destacou Silva (2011, p. 19), os estudos sobre a escravidão na região semiárida da Bahia ainda careçam de maior atenção, há estudiosos, entre os quais Poppino (1968), que citam a presença de negros fugidios no sertão, que teriam formado pequenos quilombos em suas matas, ou alguns poucos escravos que trabalharam na elementar agricultura (pois, no sertão, o que mais se desenvolvera foi a atividade pecuária). Sobre essa questão, o entendimento que se sustenta neste texto é o de que a maior concentração de escravos no município de Feira de Santana deva ter se dado mais a partir do século XIX, com a plantação de lavouras de algodão e também com o recebimento de muitos ex-escravos que vieram trabalhar na região. Nesse período do final do século XIX e início do XX, há também que se considerar que muitos brancos e mestiços migraram para Feira de Santana para trabalharem na cidade, que a essa altura, já delineava sua forte vocação: a de ser um polo comercial, consolidando-se como "um empório do sertão", denominação atribuída comumente por jornalistas da época, conforme informa Oliveira (2000, p. 9).

Tendo o município prosperado muito, saindo da condição de uma "singela" feira de gado e transformando-se numa cidade com características desenvolvidas/modernizadas, houve muitas alterações no seu quadro populacional, principalmente a partir das primeiras cinco décadas do século XX. Poppino (1968), dado

3 Acredita-se ser mais seguro afirmar que, pelo menos até o início do século XVII, os vaqueiros tinham uma origem indígena, pelo fato de os índios terem mais habilidades em embrenharem-se pelos caminhos do sertão. Após esse período, os vaqueiros deveriam ser mestiços, com ascendência indígena, negra ou mesmo branca.

4 No final do século XVII, o português João Peixoto Viegas se estabeleceu no atual distrito de Maria Quitéria (GALVÃO, 1982).

o rápido desenvolvimento urbano do Município, chama atenção para o fato de Feira de Santana ter prosperado, em menos de um século e meio; crescimento acelerado que, por sua vez, teria acontecido também em outras cidades do Brasil a partir do século XIX. Mas, para o autor, o caso de Feira de Santana é singular, dado que a Bahia estava justamente em declínio, em comparação ao que houve nos áureos tempos do período colonial. Para o autor, são notáveis as forças políticas, econômicas e sociais que impulsionaram o extraordinário desenvolvimento em Feira de Santana.

Essa característica atrativa de Feira de Santana intensificou-se a partir das primeiras décadas do século XX, tendo atraído, inclusive, muitos migrantes nordestinos que se instalaram na cidade para atuarem no comércio, como aliás, pode ser aferido pelos nomes de pioneiras lojas do comércio local, como A Cearense, Sobral, entre outras.

No século XXI, Feira de Santana continua a atrair pessoas, não só da circunvizinhança, mas de outros estados, para trabalharem em suas indústrias. Após a Segunda Guerra Mundial, a cidade congregou mais motivos para aumentar o desenvolvimento de indústrias na região, principalmente pelo aumento comercial, entre os anos de 1940 e 1950, em virtude do crescimento da população, do progresso dos transportes e da dificuldade de importação de produtos (POPPINO, 1968). Se até os anos 1940, a indústria era incipiente, apenas com aproveitamento de carne e de gêneros alimentícios, hoje, conta com vários outros produtos, inclusive, com fábricas multinacionais, como a da Nestlé. O município saiu da condição de comunidade rural para a de centro comercial e industrial de grande importância no estado da Bahia e do Brasil.[5]

Se as características da cidade, até o final do século XIX, moldavam-se ao que se tinha no restante do Brasil, isto é, um país com características eminentemente rurais, com uma grande parte da população dedicando-se a atividades agrícolas, e prioritariamente, residindo na zona rural[6], a partir de 1940, a situação se altera.

5 A partir da década de 1970, o desenvolvimento industrial da cidade foi impulsionado com a criação do Centro das Indústrias de Feira de Santana (CIFS) e do Centro Industrial do Subaé (CIS), que atraíram ainda mais migrantes de todas as regiões para a cidade, que vislumbravam possibilidades de trabalho e ofertas de serviços. Antes do efetivo desenvolvimento industrial em Feira de Santana, Moreira (1986) destaca a forte presença de muitos migrantes ao trabalho como ambulantes e feirantes, na década de 1970, malgrado o sucesso imediato da implantação do CIS.

6 Segundo Poppino (1968, p. 188-189), o recenseamento de 1872 mostrou que quase 90% da população adulta se constituíam de agricultores: "de um total de 33 mil habitantes, acima de 16 anos aproximadamente 29 mil". No Censo de 1920, novamente a população foi categorizada conforme as profissões e predominavam atividades agrícolas. Em 1940, 78% dos habitantes do município se incluíam na população rural. Já em 1950, esse número cairia para 68%. A partir dessa década, com o crescimento das atividades comerciais e industriais, a população rural diminuiria ainda mais.

Sobre o impacto da industrialização e do comércio no aumento demográfico no município, a Tabela 1 mostra a predominância urbana da população feirense nas últimas décadas, em contraste com havia em décadas passadas.

ANOS	POPULAÇÃO RESIDENTE					
	TOTAL	(%) [1]	URBANA	(%) [1]	RURAL	(%) [1]
1940	83.268	–	19.660	–	63.608	–
1950	107.205	28,75	34.277	74,35	72.928	14,65
1960	141.757	32,23	69.884	103,88	71.873	–1,44
1970	187.290	32,12	131.720	88,48	55.570	–22,68
1980	291.504	55,65	233.905	77,58	57.599	3,65
1991	406.447	39,43	348.973	49,20	56.875	–1,26
2000	480.949	18,33	431.730	23,71	49.219	–13,46
2010	556.642	15,74	510.637	18,28	46.007	–6,53

[1] Variação percentual com o período imediatamente anterior.

Tabela 1 – Crescimento absoluto e relativo da população urbana e rural residente no município de Feira de Santana/BA, 1940 – 2010. Fonte: Anuário Estatístico de Feira de Santana – 2012.

Na Tabela 1, merece destaque o crescimento relativo da população urbana, apresentado na quinta coluna, com ápice na década de 1960 (103,88% em relação à década de 1950). Por outro lado, Oliveira (2000), ao estudar o período de 1883 a 1937, identificou e analisou os processos de destruição da ordem rural em Feira de Santana. A hipótese central foi a de que, durante os anos finais do século XIX e as três primeiras décadas do XX, houve profundas mudanças na cidade e que "essas transformações, articuladas entre si, produziram novidades em termos de modelos de sociabilidade, gerando um novo padrão de comportamento público e uma nova 'urbe'". (OLIVEIRA, 2009, p. 17). Na sua interpretação, essas mudanças estariam inseridas no ideal republicano e estariam em consonância com as ideias iluministas de trazer progresso, pautadas na ciência e na razão e, não raro, na crença de que uma sociedade sem mestiçagem e com padrões urbanos seria mais propícia para ser "evoluída". No que diz respeito à sociedade feirense, o autor assim pronuncia-se:

> A construção de um novo comportamento público foi feita em meio a vários conflitos, destacadamente contra as heranças da cultura negra, dos vaqueiros e de outras formas de ação que lembrassem o passado pastoril da cidade. Feira de Santana então é transformada em uma verdadeira arena de conflitos, na acepção de Henri Lefebvre, na qual o centro era a escolha

> das melhores maneiras de organização da população no espaço público, com a exclusão daqueles setores indesejáveis às novas formas de sociabilidade. (OLIVEIRA, 2000, p. 18)

Nota-se uma preocupação voltada para apagar as raízes rurais da cidade, estando em pauta o desejo de tornar Feira de Santana uma "cidade grande", ficando atrás apenas da capital do estado, ou, como destacou Oliveira (2000), o desejo era o de tornar Feira de Santana a "Petrópolis baiana":

> Em uma cidade construída no interior da Bahia, com sólidas bases rurais, certamente a chegada de tais novidades provocaram conflitos, uma vez que ficaram em choque as duas principais características de Feira de Santana: de um lado o passado rural e do outro o fortíssimo incremento do comércio e o conseqüente desenvolvimento urbano. (OLIVEIRA, 2000, p. 25)

O autor continua a discorrer sobre o conflito rural e urbano em Feira de Santana num estudo posterior (OLIVEIRA, 2011). Discute, por exemplo, uma matéria jornalística, datada de 4 de maio de 1929, em que a chamada elite culta feirense manifesta o seu desejo de que a Feira de Santana se assemelhasse, cada vez mais, à capital:

> Rejubilem-se os caminhões. Carroças na capital só até 31 de dezembro. Uma lei do Conselho Municipal da cidade do Salvador proíbe o trafego de carroças a partir de 1º de janeiro de 1930. Os pobres muares que subiam o Taboão, o Caminho Novo, a Água Brusca, Montanha, Santa Tereza e outras ladeiras da capital, às vezes sobre desproporcionadas cargas, irão resfolegar do Ano Bom em diante e os carroceiros que vão desde já procurando outra vida; porque, felizmente para eles vão ficar esquecidas suas barbaridades revoltantes, como ficaram olvidadas as dos aguadeiros, que já há muitos anos deram o fora do perímetro urbano da capital. Quando será que também nos veremos livres dos daqui? (Folha do Norte de 04 /05/1929 apud OLIVEIRA, 2011, p. 34-35).

Correlacionando à formação da língua falada em Feira de Santana, presume-se que os usos linguísticos característicos dos falares rurais/populares teriam sofrido, até de maneira inconsciente, uma forte campanha para serem banidos do espaço urbano. Considera-se que, nesse contexto sócio-histórico, seriam muito mais "adequados" os falares que mais se aproximassem de usos urbanos/letrados, não cabendo, assim, usos linguísticos estigmatizados socialmente e rotulados como típicos da fala rural ou próxima dessa, a saber, a de pessoas com baixa ou nula escolaridade (historicamente, no Brasil, essas habitaram o espaço rural/interior). Dado que os inúmeros migrantes que se radicaram em Feira de Santana tinham esse perfil, foi ainda mais imperativo considerar a questão da migração na constituição da amostra de fala vernácula feirense.

2. CRITÉRIOS DE CONSTITUIÇÃO DA AMOSTRA

O projeto *A Língua Portuguesa no Semiárido Baiano* sempre procurou utilizar critérios socio-histórico e demográficos na etapa de constituição de suas amostras e, na sua terceira fase, investiu-se intensamente nessa vertente socio-histórica, tornando-se imprescindível acolher os postulados dos estudos da área das ciências sociais e humanas. Assim se procedeu por se intentar trazer elementos para melhor discutir a formação e a consolidação da língua portuguesa na comunidade de fala em foco, além de oferecer embasamentos para melhor se interpretar os resultados acerca da realidade sociolinguística feirense.

Em decorrência dos conhecimentos sobre aspectos da socio-história e da demografia da comunidade de fala, consideraram-se, na constituição da amostra, além de critérios comumente empregados nas pesquisas sociolinguísticas (como *faixa etária*, *sexo* e *escolaridade* do informante), a *relação do informante com a migração*. Por conseguinte, a amostra urbana de Feira de Santana foi dividida em subamostras, sendo considerado fatores sobre o informante ser feirense, filho de feirenses, feirense filho de migrantes e migrante. No Quadro 1, são apresentados os critérios socioculturais utilizados para a seleção dos informantes e, por conseguinte, para a constituição da amostra.

FATORES SOCIOCULTURAIS	
GÊNERO	Masculino Feminino
FAIXA ETÁRIA	Faixa 1 (25-35 anos) Faixa 2 (45- 55 anos) Faixa 3 (acima de 65 anos)
RELAÇÃO COM A MIGRAÇÃO	Feirenses filhos de feirenses Feirenses filhos de migrantes Migrantes
ESCOLARIDADE	Baixa ou inexistente (analfabetos ou semialfabetizados) Ensino médio completo Ensino superior completo com ou sem pós-graduação

Quadro 1 – Fatores socioculturais utilizados na constituição da amostra do Projeto de Pesquisa *A Língua Portuguesa no Semiárido Baiano – Fase 3*.

No acervo, há entrevistas com informantes da zona rural do município, especificamente no distrito da Matinha, gravadas durante a segunda fase do

do projeto. Algumas mais foram gravadas durante a Fase 3, a fim de torná-las compatíveis com a nova faixa etária dos informantes, alterada no ano de 2008. Consequentemente, é possível comparar os resultados da zona urbana com os da zona rural do município. No que concerne à comunidade de fala de Feira de Santana, há 72 entrevistas gravadas e transcritas, assim distribuídas:

- 48 com informantes analfabetos ou pouco escolarizados[7], sendo 12 informantes da zona rural e 36 da sede do município. No caso destes últimos, 12 *filhos de feirenses*, 12 *filhos de migrantes* e 12 *migrantes*. Os da zona rural são nascidos no município e os seus pais, na maioria dos casos, também;
- 12 informantes com ensino superior completo e/ ou com pós-graduação (todos informantes da sede do município, nascidos na própria cidade);
- 12 informantes com Ensino Médio completo, sendo feirenses e filhos de feirenses.

A amostra possibilita uma análise contextualizada acerca do binômio variação/mudança, abrangendo aspectos marcantes na sócio-história da comunidade de fala, a exemplo dos contatos interdialetais e dos tardios processos de escolarização e de urbanização do município. As entrevistas também possibilitam a realização de estudos que investiguem a configuração atual dos dois grandes polos sociolinguísticos do português brasileiro – PB, isto é, o que, na formulação teórica de Lucchesi (1994; 2001; et al.), denominam-se *norma culta* e *norma popular*. Os estudos com os dados linguísticos de Feira de Santana podem lançar "luzes" sobre a polêmica formação do português brasileiro, além de reunir elementos que permitam investigar quais as consequências do estreitamento das redes sociais empreendido no Brasil, a partir da intensificação dos processos de urbanização, bem como o da democratização de acesso ao ensino. Em outras palavras, possibilita a realização de pesquisas sobre a caracterização sociolinguística atual, tomando como base a comunidade de fala feirense.

Sustenta-se a hipótese de que a formação do português brasileiro foi marcada por determinadas condições sociais que o fizeram ser diferente do português europeu, a exemplo do intenso contato entre povos e línguas e da tardia implantação

7 A intenção era gravar apenas informantes que tivessem estudado por até quatro anos, porém, em vista da dificuldade de serem encontradas pessoas com essa característica (principalmente na faixa 1), foram gravadas entrevistas com informantes que estiveram na escola por mais tempo, sendo que alguns estavam concluindo o Ensino Fundamental, mas no supletivo, em que se estudam duas séries em um ano. Nesse sentido, considerando as deficiências do ensino que frequentaram e, principalmente, que as suas atividades profissionais não lhes proporcionavam maior contato com o letramento, julgamos que a característica popular da sua norma linguística ficou preservada.

dos processos educacionais e urbanísticos brasileiros, associados a um perverso sistema de discriminação racial e exclusão social. Tais condições repercutiram significativamente na estrutura da língua portuguesa, fazendo com que houvesse uma bipolarização de normas linguísticas no Brasil, com um polo que abriga as variedades cultas, mais próximas da norma padrão, e outro que abriga as variedades populares, marcadas por um processo de extrema redução da morfologia flexional. Entendemos igualmente que, com as profundas e contínuas mudanças ocorridas no Brasil, a partir do século XX, tais normas podem estar em processo de entrecruzamentos, influenciando-se mutuamente (LUCCHESI, 2001).

A subamostra com informantes de ensino médio completo não tem sido muito utilizada no âmbito do projeto, constituindo-se, na verdade, uma "subamostra-controle", pois é problemático considerar o falar desses informantes pela seguinte razão: qual a norma linguística que esses informantes utilizam? Isto é, num *continuum*, em que mais se aproximam esses falares, do polo culto ou do popular? Acredita-se que, levando em consideração a história da escolarização da população brasileira, bem como outros dados da sócio-história do português brasileiro, seja necessário distinguir os dados linguísticos provenientes de falantes com ensino médio dos tempos atuais (após as "facilidades" de acesso ao ensino formal) da fala de informantes com esse perfil de décadas passadas. Explica-se: não se considera apropriado colocar como equivalentes dados linguísticos provenientes da fala de informantes da faixa III com a de informantes da faixa II e I, em caso de ambos terem escolaridade secundária. Isso se justifica por entendermos que há diferenças qualitativas na escolarização de um informante idoso que estudou o ensino médio, numa época em que a escola era elitizada e elitista, da de uma pessoa adulta (com, por exemplo, 30 anos) com apenas o ensino médio e, muitas vezes, concluído pelo supletivo. Na amostra, inclusive, há um informante idoso que só estudou até o ensino médio, mas que ocupou o cargo de gerente de um banco. Assim, resultados de análises que consideram a fala de informantes como "semicultos" (nível intermediário de escolarização) podem conter enviesamentos quantitativos e qualitativos, na medida em que ora pode ter traços linguísticos do vernáculo brasileiro (ou da norma popular do português brasileiro), ora do falar culto. Portanto, se o objetivo for apreender o vernáculo brasileiro (livre de normatizações), é preciso considerar esse aspecto.

No caso da subamostra com os migrantes, há também uma ressalva a ser feita. Primeiramente, como os estudos historiográficos e demográficos apontavam que a grande parte dos migrantes que vieram para o município era formada basicamente por pessoas com baixa ou nula escolaridade, foi apenas considerado esse aspecto dos informantes usuários da norma popular (utilizada por informantes com baixa ou nula escolaridade). Em etapas futuras, poderão ser constituídas amostras com migrantes com alta escolarização. Já no caso dos informantes com

ensino médio ou superior, foram escolhidos feirenses cujos pais também fossem feirenses, como, aliás, foi adotado na comunidade rural da Matinha.

O enfoque principal do projeto é justamente apreender as características do vernáculo popular feirense, comparando-o com o falado em outras regiões do estado da Bahia e contrastando-o com o falado por pessoas com escolarização máxima, nascidas e residentes no município. O intuito de ter sido constituída a amostragem assim bipolarizada foi o de permitir a comparação entre a norma popular e a culta, com base na visão bipolarizada da realidade sociolinguística brasileira (LUCCHESI, 1994; 2001; et al.).

Os estudos já realizados permitem asseverar que, a depender do fenômeno analisado, a situação sociolinguística bipolarizada do português brasileiro mantém-se na fala feirense, notadamente quando o fenômeno linguístico é estigmatizado socialmente. Araujo (2012), por exemplo, ao investigar o uso variável da concordância verbal com a primeira pessoa do plural tomando como amostra a subamostra rural (12 entrevistas), encontrou uma variação bem estruturada. Dos 44 dados levantados, 18, que correspondem a 40,9%, foram com a variante zero (ou sem marcas de plural). A autora ressalvou que o índice de não realização dos morfemas "-mos ~ -mo ~ -emo" só não deve ter sido maior devido ao amplo uso da forma pronominal "a gente" acompanhada de formas verbais sem marcas explícitas de plural. Ao comparar os resultados encontrados com os obtidos a partir do levantamento de dados na amostra do projeto português Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN http://www.clul.ul.pt/sectores/variacao/cordialsin/projecto_cordialsin_corpus.php), a autora encontrou, na amostra portuguesa, também com informantes da zona rural com baixa ou nula escolarização, um resultado diferente: apenas uma ocorrência sem morfema de plural dos 128 dados levantados (99,2%). Tais resultados levaram a autora a afirmar que:

> Embora o contexto sócio-histórico do período da formação da realidade sociolinguística brasileira não tenha dado ensejo à formação de línguas crioulas prototípicas, não se pode deixar de ver, ainda hoje, no português popular brasileiro, influências do intenso contato entre línguas, bem como das situações de exclusão social sofridas pela população afrodescendente ao longo da história. (ARAUJO, 2012, p. 91)

Já no estudo de Araujo (2014), ao comparar dados levantados na fala de 36 informantes com baixa ou nula escolarização (feirenses filhos de feirenses, feirenses filhos de migrantes e feirenses da zona rural) com os levantados na fala dos 12 informantes com ensino superior, focalizando a variação na concordância verbal com a terceira pessoa do plural, fora identificado um quadro bem polarizado.

NORMA CULTA			NORMA POPULAR		
	Aplicação/Total	%		Aplicação/Total I	%
Variante padrão	619/659	93,9%	Variante padrão	321/1310	24,5%
Variante não padrão	40/659	6,1%	Variante não padrão	989/1310	75,5%

Tabela 2 – Distribuição geral dos resultados nas subamostras pesquisadas no estudo de Araujo (2014).

Observa-se, na Tabela 2, que, enquanto os resultados com os informantes com alta escolaridade indicam haver uma variação marginal e residual, com os informantes analfabetos ou com baixa escolarização a variação é bem marcante, com amplo ou zero uso da variante não padrão. O trabalho de Santos (2014), que analisa a variação entre o futuro do pretérito – FP e o pretérito imperfeito – PI em contextos *irrealis* no português culto e popular falado em Feira de Santana, demonstra que os falantes menos escolarizados favorecem o uso de PI e esta variante é menos usada pelos falantes mais escolarizados, que optam pela variante FP, que parece ter maior prestígio social.

Tais resultados ratificam a polarização da realidade sociolinguística brasileira, destacando a pertinência do conceito de *norma linguística* no âmbito da teoria da Sociolinguística Variacionista, conforme trabalhado por Lucchesi (1994, 2001, 2002 e 2006)[8], que define dois parâmetros para a realidade linguística brasileira, o da *norma culta* e o da *norma popular*, chamando a atenção que as contínuas mudanças ocorridas na sociedade brasileira a partir das primeiras décadas do século XX podem diminuir o abismo que, historicamente, separava a fala da classe social baixa da fala da elite:

> A polarização lingüística do Brasil não é, porém, estanque, podendo-se detectar influxos que interligam os dois subsistemas distintos, sobretudo a partir das primeiras décadas do Século XX, quando se inicia o vigoroso e profundo processo de industrialização e urbanização do país, que dinamizou a reprodução da cultura e democratizou as relações sociais, sem conseguir entretanto alterar o quadro de profundas desigualdades sociais que ainda entravam o verdadeiro desenvolvimento do país. As contradições da realidade social refletem-se no plano das normas lingüísticas, pois, ao tempo em que se observa, no plano objetivos dos padrões coletivos de comportamento verbal, uma tendência ao nivelamento das duas normas lingüísticas brasileiras, no plano subjetivo da avaliação das variantes lingüísticas, o estigma ainda recais pesadamente sobre as variantes mais características da norma popular, fortalecendo-se, a cada dia – inclusive com a força

8 Uma norma linguística é definida não por pessoas que falam de maneira igual, mas que compartilham idêntico sistema de avaliação subjetiva das variantes linguísticas (WEINREICH; LABOV; HERZOG, 2006).

> dos meios de comunicação de massa – um preconceito que, sem fundamento lingüístico (cf. Bagno, 1999), nada mais é do que a crua manifestação da discriminação econômica e da ideologia da exclusão social. (LUCCHESI, 2002, p. 87-88)

Por outro lado, os estudos já realizados com a amostra de Feira de Santana, que envolvem fenômenos linguísticos não marcados socialmente, têm evidenciado a aproximação dos dois polos sociolinguísticos, aludida por Lucchesi (2002), indo também ao encontro da visão de dois *continua* (o de urbanização e o de letramento), postulada por Bortoni-Ricardo (2005, 2008).

> Todo falante do português do Brasil situa-se em um ponto determinado desse contínuo, mas pode movimentar-se em direção a qualquer dos pólos, dependendo de sua rede de relações sociais, sua inserção em práticas sociais letradas e participação no sistema de produção, bem como seu gênero, faixa etária e outros componentes de sua identidade social. O contínuo de urbanização permite ainda distinguir regras variáveis graduais, presentes ao longo de todo o contínuo, e regras descontínuas, características do repertório das populações situadas no pólo rural e na zona urbana (BORTONI-RICARDO et al, 2008, p.231).

O estudo realizado por Santana (2014) sobre a realização do objeto direto e do indireto de terceira pessoa no português falado em Feira de Santana, demonstrou que não há diferenças significativas de uso entre os falantes com baixa ou nula escolaridade e os com curso superior, isto é, entre os representantes da norma popular e os da culta. O clítico "lhe", por exemplo, atuando como complemento direto e indireto referente à 2ª pessoa, aparece com bastante frequência no *corpus* analisado, tanto no português popular quanto no culto. O estudo também detectou um amplo uso do objeto nulo em todo o *corpus*, independentemente da escolaridade do falante. Esses resultados, associados a outros, sugerem que essa variante não é interpretada como estigmatizada pelos usuários do português popular e culto feirenses.

Feitas essas observações acerca, principalmente, dos critérios adotados para a constituição da amostra da comunidade de fala de Feira de Santana, no que tange à temática da polarização sociolinguística, discorremos sobre como foi realizado o controle das outras variáveis socioculturais na fase de coleta da amostra. Quanto ao critério sexo do informante, trabalhou-se com ambos, ignorando a sua orientação sexual. Apenas um informante culto, do sexo masculino e da faixa etária I revelou, durante a entrevista, ser homossexual e, nesse caso, foi contabilizado como do sexo masculino. Essa é uma questão que merece ser refletida com maior acuidade na constituição de futuros acervos linguísticos.

Quanto à variável faixa etária, sendo importantíssima por permitir as projeções históricas acerca de fenômenos variáveis, pondo em destaque o binômio

variação e mudança, foi adotado o critério de faixas descontínuas, respeitando-se o intervalo de dez anos entre elas, por julgarmos que assim as faixas seriam, de fato, representativas dos três grupos etários considerados: o jovem, o mediano e o idoso. As faixas etárias englobam informantes com idades em que subjazem, de fato, aspectos sociocomportamentais de indivíduos jovens, adultos e idosos, o que não ocorreria se as faixas fossem contínuas[9]. O intervalo das faixas etárias adotado foi o mesmo do utilizado no âmbito do projeto "Vertentes do português popular do Estado da Bahia", de maneira que as amostras tornam-se idealmente intercomparáveis. A variável foi controlada da seguinte forma: Faixa 1 – informantes entre 25 a 35 anos; Faixa 2 – informantes entre 45 a 55 anos; Faixa 3 – informantes a partir de 65 anos (Quadro 1).

Os dados sócio-históricos do município foram fundamentais não apenas para a definição dos critérios de seleção dos informantes, mas também durante a realização de estudos variacionistas, orientando a formulação das hipóteses e a interpretação dos resultados. A título de exemplo, na pesquisa realizada por Araujo (2014) sobre a variação na concordância verbal com a terceira pessoa do plural, ao identificar entre os informantes cultos um favorecimento do uso padrão apenas entre informantes da faixa III (média de idade de 67,7 anos), e considerando a hipótese clássica acerca da variável faixa etária, a autora ponderou que uma possível explicação poderia ser buscada no fato de apenas esses informantes terem formado o seu vernáculo antes da intensificação dos processos de contato dialetais no município, bem como dos processos de democratização de acesso ao ensino.[10]

3. CRITÉRIOS DA AMOSTRAGEM

Quanto ao tamanho da amostragem, ou melhor, ao número de informantes que compõem a amostra da comunidade de fala de Feira de Santana, o total é

9 Muitos projetos desenvolvidos no Brasil não ponderaram essa questão, de modo que as idades limítrofes das faixas etárias estão muito próximas. O projeto NURC, por exemplo, cujas faixas etárias estabelecidas foram I (25 a 35 anos), II (36 a 55 anos) e III (56 em adiante) ilustra esse fato. Em casos como esse, pergunta-se: quais as possíveis diferenças externadas na fala de um indivíduo com 35 e 36 anos, ou ainda, de 55 ou 56 anos? Um indivíduo com 56 anos já poderia ser considerado idoso nos dias atuais?

10 Sobre essa questão, devem ser considerados estudos sobre a urbanização e escolarização, notadamente sobre a demografia histórica feirense, a exemplo do fato de, em 1950, o município ter apenas 32% de sua população residente no perímetro urbano e, nas décadas seguintes, esses percentuais terem se invertido, graças à migração de uma grande leva de pessoas vindas do campo e de cidades menores (FREITAS, 1998, p. 125).

de 60, na zona urbana, e de 72, incluindo também a subamostra da zona rural. Assim, há 12 informantes em cada uma das subamostras, sendo dois em cada uma das células (considerando o sexo e a idade do informante). Nos Quadros 2-7 a constituição e a amostragem das subamostras são detalhadas.

	Masculino	**Masculino**	**Feminino**	**Feminino**
Faixa I (25 a 35 anos)	26 anos Pedreiro 5ª série	35 anos Vigilante 4ª série	33 anos Diarista 5ª série	31 anos Doméstica 5ª série
Faixa II (45 a 55 anos)	50 anos Entregador de recibos da Coelba 2ª série	45 anos Pintor de parede 5ª série	54 anos Empregada doméstica 4ª série	50 anos Dona de bar 2ª série
Faixa III (mais de 65)	72 anos Pedreiro 2ª série	80 anos Pedreiro/carpinteiro Analfabeto	70 anos Dona de casa 3ª série	76 anos Merendeira 3ª série

Quadro 2 – Apresentação dos informantes da norma popular urbana (feirenses filhos de feirenses).

	Masculino	**Masculino**	**Feminino**	**Feminino**
Faixa I (25 a 35 anos)	32 anos Comerciante 3ª série	35 anos Pintor de parede 3ª série	28 anos Ajudante de cozinha 4ª série	27 anos Dona de casa 4ª série
Faixa II (45 a 55 anos)	53 anos Representante comercial 3ª série	45 anos Catador de papel 2ª série	48 anos Dona de casa 8ª série	48 anos Empregada doméstica 1ª série
Faixa III (mais de 65)	66 anos Aposentado (Serviços gerais) Analfabeto	82 anos Lavrador 3ª série	69 anos Aposentada (Lavradora) 3ª série	66 anos Empregada doméstica 3ª série

Quadro 3 – Apresentação dos informantes da norma popular (feirenses filhos de migrantes).

	Masculino	Masculino	Feminino	Feminino
Faixa I (25 a 35 anos)	29 anos Motorista de ônibus 7ª série *Alagoinha –Pernambuco*	30 anos Vigilante 5ª série *Milagres – Bahia*	33 anos Comerciária 8ª série *Bonfim de Feira – Bahia*	26 anos Dona de casa 7ª série *Pé de Serra –Bahia*
Faixa II (45 a 55 anos)	47 anos Motorista de ônibus 6ª série *Tanquinho – Bahia*	49 anos Aposentado (pintor) 5ª série *Riachão de Jacuípe – Bahia*	45 anos Diarista 2ª série *Serra Preta –Bahia*	55 anos Costureira 4ª série *Santa Bárbara – Bahia*
Faixa III (mais de 65)	84 anos Aposentado (Petroleiro) 4ª série *Serrinha – Bahia*	82 anos Aposentado (Pedreiro) Analfabeto *Serrinha – Bahia*	66 anos Dona de casa 3ª série *Campina Grande – Paraíba*	75 anos Dona de casa 4ª série *Jaguaraí –Minas Gerais*

Quadro 4 – Apresentação dos informantes da norma popular (migrantes).

	Masculino	Masculino	Feminino	Feminino
Faixa I (25 a 35 anos)	31 anos Pedreiro 4ª série	35 anos Serigrafista 5ª série	28 anos Lavradora 4ª série	32 anos Lavradora 2ª série
Faixa II (45 a 55 anos)	52 anos Pedreiro 4ª série	48 anos Vaqueiro 2ª série	55 anos Lavradora Analfabeta	55 anos Lavradora 2ª série
Faixa III (mais de 65)	74 anos Lavrador 3ª série	74 anos Lavrador e comerciante 3ª série	68 anos Lavradora Analfabeta	77 anos Lavradora Analfabeta

Quadro 5 – Apresentação dos informantes da norma rural (feirenses filhos de feirenses).

	Masculino	Masculino	Feminino	Feminino
Faixa I (25 a 35 anos)	33 anos Contador	25 anos Historiador	26 anos Enfermeira	30 anos Administradora
Faixa II (45 a 55 anos)	53 anos Engenheiro civil e professor universitário	56 anos Químico/ professor Ensino Médio	48 anos Professora universitária	45 anos Professora e mestranda em Desenho
Faixa III (mais de 65)	69 anos Economista e contador	67 Arquiteto e artista plástico	68 anos Jornalista	67 anos Pedagoga

Quadro 6 – Apresentação dos informantes da norma culta (feirenses filhos de feirenses).

	Masculino	Masculino	Feminino	Feminino
Faixa I (25 a 35 anos)	26 anos Mecânico	35 anos Auxiliar de escritório	33 anos Técnica em Enfermagem	33 anos Atendente em escritório
Faixa II (45 a 55 anos)	47 anos Técnico em Telefonia	49 anos Mecânico industrial	45 anos Agente penitenciária	49 anos Funcionária pública
Faixa III (mais de 65)	68 anos Aposentado (Gerente de Banco)	81 anos Aposentado (Comerciário)	69 anos Professora primária	73 anos Professora primária

Quadro 7 – Apresentação dos informantes com Ensino Médio (feirenses filhos de feirenses).

Seguindo a orientação comumente empregada nos manuais de Sociolinguística, o número de informantes deveria ser maior, já que a recomendação é que seja feita uma análise combinatória, multiplicando o número de fatores das variáveis sociais consideradas (OLIVEIRA E SILVA, 2003). Se fosse adotada a indicação clássica, teria se procedido da seguinte maneira: 2 sexos (H, M) x 3 faixas etárias (I, II, III) x 3 níveis de escolaridade (baixa ou nenhuma, ensino médio, ensino superior) x 3 tipos de relação com a migração (feirense filho de feirense, feirense filho de migrante, migrante) x 2 zonas (rural e urbana) = 108. E assim, também seguindo a orientação clássica (LABOV, 2008), multiplicaria esse número (que já é elevado) por 5 (já que deveria haver 5 informantes em cada célula). Ou seja, 108 x 5 = 540 informantes. Obviamente, esse número, embora ideal, é inviável para a concretização das pesquisas sociolinguísticas, que, como todos sabem, persistem em meio a tantos obstáculos (falta de financiamentos, dificuldades em se conseguir informantes com certos perfis, perda de entrevistas por problemas técnicos, etc.). Logo, se fosse adotada a recomendação ideal, a finalização da constituição da amostra estaria longe de ser encerrada, já que, se com o número de 72 informantes foram necessários quase cinco anos de árduo trabalho, com a supervisão de quatro pesquisadores e mais de uma dezena estudantes. Imaginem se fossem 540 entrevistas!

Não obstante a essas ponderações, frisamos que não há mais entrevistas na amostra não apenas pela onerosidade de tempo e de custos, mas principalmente, devido aos objetivos do projeto. Os conhecimentos norteados pela história externa da língua falada em Feira de Santana indicaram que era imperioso haver células vazias na amostra. Explica-se: (I) na zona rural do município, praticamente inexistem pessoas com escolaridade superior; (II) não fazia sentido gravar migrantes com escolarização alta, uma vez que os estudos apontavam que a reação negativa dos feirenses dizia respeito aos migrantes com baixa ou nula escolaridade e que foi esse o perfil dos migrantes que massivamente radicaram-se em Feira de Santana.

Portanto, embora haja ajustes a serem feitos na amostra da fala feirense – como, por exemplo, aumentar a amostragem referente à norma culta (a fim de torná-la mais compatível com a taxa de pessoas cultas do município) – acredita-se que a amostragem aleatória deu conta de representar a comunidade de fala de Feira de Santana, oferecendo um banco de dados que permita que diversas pesquisas linguísticas sejam realizadas, a fim de se investigar diversos fenômenos, não excluindo a realização de pesquisas sociológicas e antropológicas.

CONSIDERAÇÕES FINAIS

Há muitas dificuldades na formação de *corpora* orais, no entanto, buscou-se minimizar ao máximo os problemas, deixando claros os critérios que nortearam a constituição da amostra. Há limitações pelo fato de os pesquisadores terem se concentrado apenas na entrevista tipo Diálogo entre Informante e Documentador (DID); esse tipo de entrevista não é muito eficiente na captação de alguns fenômenos, como, por exemplo, a segunda pessoa. No entanto, tentou-se minimizar esse problema realizando, ao final da entrevista, uma pergunta direta ao informante sobre o uso de "tu/você". Por enquanto, só com este fenômeno (tu/você) observaram-se poucas ocorrências. Para diversos outros fenômenos, a amostra tem se mostrado bastante representativa, a exemplo da variação entre futuro simples e perifrástico, a variação entre o futuro do pretérito e do pretérito imperfeito, variação na colocação pronominal, variação nas concordâncias, topicalização, entre outros.

A amostra pode ainda ser complementada, não só com a ampliação das entrevistas do tipo DID, mas também com a gravação de conferências, aulas e também de comunidades de prática, a exemplo dos professores universitários (presentes na amostra dos representantes do culto). No entanto, julga-se que o material é bastante interessante para se trabalhar com o contínuo rural/urbano e iletrado/letrado, além de análises que levem em consideração diferentes questões sociais, a exemplo da questão da migração, do contato com o rural, entre outras.

REFERÊNCIAS

ALMEIDA, N. L. F. de; ZENAIDE O. N. A língua portuguesa falada no semi-árido baiano: algumas considerações. In: Múltiplos olhares sobre a língua, 1999, Maceió. *Anais do Congresso de Língua Falada e Escrita*, Maceió: UFAL, 1999.

_____. *Coleção amostras da língua falada no semiárido baiano*. Feira de Santana: Universidade Estadual de Feira de Santana/FAPESB, 2008.

_____. *A variação linguística no semiárido baiano*. Feira de Santana: UEFS Editora, 2014.

ARAUJO, S. S. F. A concordância verbal no português falado em Feira de Santana-BA: sociolinguística e sócio-história do português brasileiro. Salvador, 2014. Tese (Doutorado em Língua e Cultura) - Instituto de Letras, Universidade Federal da Bahia.

ARAUJO, S. S. F.; ARAUJO, J. M. O. A formação sócio-histórica do português do Brasil: contribuições do recôncavo baiano. *Cadernos de Letras da* UFF – Dossiê: Difusão da língua portuguesa, Niterói: n. 39, p. 95-116, 2009.

BORTONI-RICARDO, S. M.; SILVA, M. G. T.; CAXANGÁ, M. R. R.; LINS, M. V. Raízes sociolinguísticas do analfabetismo no Brasil. *Revista acolhendo a alfabetização nos países de língua portuguesa*, São Paulo: n. 04, março/agosto de 2008. p. 215-234.

FREITAS, N. B. *Urbanização em Feira de Santana: influência da industrialização 1970-1996*. Salvador, 1998. Dissertação (Mestrado em Arquitetura) – Faculdade de Arquitetura, Universidade Federal da Bahia.

GALVÃO, R. Os povoadores da região de Feira de Santana. *Sitientibus*, Feira de Santana: n.1, julho/dezembro 1982, p.25-31.

HENRIQUES, R. *Desigualdade racial no Brasil:* evolução das condições de vida na década de 90. Rio de Janeiro: IPEA, 2001.

LABOV, W. *Padrões sociolinguísticos*. Trad. M. Bagno, M. M. P. Scherre, C. R. Cardoso. São Paulo: Parábola Editorial, 2008. Título original: Sociolinguistic Patterns, 1972.

LUCCHESI, D. As duas grandes vertentes da história sociolinguística do Brasil (1500-2000). DELTA, São Paulo: v.17, n.1, 2001, p. 97-132.

_____. História do contato entre línguas no Brasil. In: LUCCHESI, D.; BAXTER, A.; RIBEIRO, (org.). *O português afro-brasileiro*. Salvador: EDUFBA, 2009. p. 41-73.

_____. Norma lingüística e realidade social. In: BAGNO, Marcos (Org.). *Linguística da norma*. São Paulo: Edições Loyola, 2002. p. 63-92.

_____. Parâmetros sociolingüísticos do português brasileiro. *Revista da* ABRALIN, v. 5, n. 1 e 2, 2006. p. 83-112.

_____. Variação e norma: elementos para uma caracterização sociolingüística do português do Brasil. *Revista Internacional de Língua Portuguesa*, n.12, 1994. p.17-28.

MOREIRA, V. D. Projeto Memória da Feira Livre de Feira de Santana. Texto nº2. A feira está morta. Viva a feira! *Sitientibus*, v. 3, n.5, 1986, p. 171-176.

OLIVEIRA E SILVA, G. Coleta de dados. In: MOLLICA, M. C.; BRAGA, M. L. (Orgs.). *Introdução à sociolinguística:* o tratamento da variação. São Paulo: Contexto, 2003. p. 117-133.

OLIVEIRA, C. F. R. M. *De empório a princesa do sertão: utopias civilizadoras em Feira de Santana (1893-1937)*. Salvador, 2000. 128 f. Dissertação (Mestrado em História) – Faculdade de Filosofia e Ciências Humanas, Universidade Federal da Bahia,

POPPINO, R. E. *Feira de Santana*. Salvador: Itapuã, 1968.

SANTANA, J. C. D. *"Todos os caminhos levam a Feira de Santana": uma viagem sociolinguística para o estudo dos pronomes objetos no português urbano falado*. Feira de Santana,

2014. Dissertação (Mestrado em Estudos Linguísticos). Departamento de Letras e Artes, Universidade Estadual de Feira de Santana.

SANTOS, A. S. *A variação entre o futuro do pretérito e o pretérito imperfeito no português falado em Feira de Santana*. Feira de Santana, 2014. Dissertação (Mestrado em Estudos linguísticos). Departamento de Letras e Artes, Universidade Estadual de Feira de Santana.

SILVA NETO, S. *Introdução ao estudo da língua portuguesa no Brasil*. 2. ed. Rio de Janeiro: INL, 1963.

SILVA, M. S. Os sertões oitocentistas na historiografia baiana: notas sobre a escravidão. In: NEVES, E. F. (Org.). *Sertões da Bahia:* Formação social, desenvolvimento econômico, evolução política e diversidade cultural. Salvador: Editora Arcádia, 2011.p. 15-50.

SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAIS DA BAHIA (SEI). Disponível em <http://www.sei.ba.gov.br>. Acesso em: 23. nov. 2011.

WEINREICH, U.; LABOV, W.; HERZOG, M. *Fundamentos empíricos para uma teoria da mudança linguística*. Trad. M. Bagno. São Paulo: Parábola Editorial, 2006. Título original: Empirical foundations for a theory of language change, 1968.

4 CAPÍTULO

A LÍNGUA FALADA EM ALAGOAS: COLETA E TRANSCRIÇÃO DOS DADOS

Elyne Giselle de Santana Lima Aguiar Vitório

INTRODUÇÃO

Ao se interessar pelo estudo da língua dentro do contexto social da comunidade de fala, a Sociolinguística Variacionista não só a vê como um fator importante na identificação e na demarcação de diferenças sociais na comunidade, como também sugere um modelo que analisa o uso variável dos fenômenos e a interferência dos condicionamentos linguísticos e sociais, proporcionando descrições mais adequadas da língua em uso pelos falantes (LABOV, 2008).

Para tanto, faz-se necessário constituir amostras sincrônicas e/ou diacrônicas da língua usada em comunidades de falas heterogêneas, tendo em vista que, para a sistematização de uma regra variável, o pesquisador sociolinguista, de acordo com Campoy e Almeida (2005), Tagliamonte (2006) e Guy e Zilles (2007), precisa definir a variável dependente e as independentes, delimitar a amostra e constituir o *corpus*, transcrever, codificar e quantificar os dados e por fim, interpretar e explicar os resultados obtidos.

Neste texto, apresentamos a metodologia e os pressupostos norteadores do trabalho de campo sociolinguístico empreendido para a constituição de uma amostra sincrônica da língua falada no estado de Alagoas. Para tanto, seguimos os pressupostos teórico-metodológicos básicos, apresentados em Guy e Zilles

(2007), e focalizamos nossa discussão em quatro momentos distintos desse processo, a saber: delimitação da comunidade estudada; constituição do *corpus* e estratificação da amostra; coleta dos dados e transcrição dos dados.[1]

1. COMUNIDADE PESQUISADA

Como não é possível compreender o processo de variação e de mudança linguística fora do contexto social de uma comunidade de fala, uma vez que, para a Sociolinguística Variacionista, a língua é uma forma de comportamento social – ou seja, a língua não é propriedade do indivíduo, mas da comunidade e, portanto, social – selecionamos a comunidade de fala alagoana e assumimos a definição de comunidade de fala proposta por Labov.

> A comunidade de fala não é definida por nenhuma concordância marcada no uso de elementos lingüísticos, mas sim pela participação num conjunto de normas compartilhadas; estas normas podem ser observadas em tipos de comportamento avaliativo explícito e pela uniformidade de padrões abstratos de variação que são invariantes no tocante a níveis particulares de uso. (LABOV, 2008, p. 150).

Para a constituição da amostra sincrônica da fala alagoana, partimos do pressuposto de que "[...] existe um conjunto uniforme de atitudes frente à linguagem que são compartilhadas por quase todos os membros da comunidade de fala, seja no uso de uma forma estigmatizada ou prestigiada da língua em questão" (LABOV, 2008, p. 176), delimitando assim, a que tipo de comunidade de fala pertence determinado indivíduo.

Conhecido como Paraíso das Águas, o estado de Alagoas está situado à leste da região Nordeste, fazendo fronteiras com o Oceano Atlântico, divisa ao Norte e ao Noroeste com o estado de Pernambuco, ao Sul com o estado de Sergipe e ao Sudeste com o estado da Bahia. Ocupa uma área de 27.779.343 km², sendo considerado um dos menores estados do Brasil, mais extenso apenas do que Sergipe. O estado de Alagoas é formado por 102 municípios e apresenta as seguintes características: o relevo é composto por planície litorânea, planalto ao norte e depressão ao centro, tendo como ponto mais elevado a serra Santa Cruz com 844 m; a vegetação é formada por floresta tropical, mangues litorâneos e caatinga; seu clima se caracteriza por ser tropical na costa e semi-árido no interior; possui o São Francisco, o Mundaú e o Paraíba do Meio como os principais rios.

---

1 A coleta de dados subsidiou a pesquisa *Ter/haver existenciais na fala alagoana: variação estável ou mudança em progresso?* (VITORIO, 2012).

As cidades alagoanas mais populosas, de acordo com o Censo 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE), são: Maceió (capital de Alagoas), Arapiraca, Palmeira dos Índios, Rio Largo, União dos Palmares, Penedo, São Miguel dos Campos, Coruripe, Campo Alegre e Delmiro Gouveia, sendo Maceió, Maragogi, Japaratinga, Barra de São Miguel, Piaçabuçu, Marechal Deodoro e Penedo os destinos turísticos mais procurados.

A população alagoana é composta por 3.120.494 habitantes, que segundo os dados da Pesquisa Nacional por Amostras de Domicílios (PNAD) – 2009, estão distribuídos entre 1.511.767 homens e 1.608.727 mulheres, apontando a predominância de habitantes do sexo feminino (Gráfico 1).

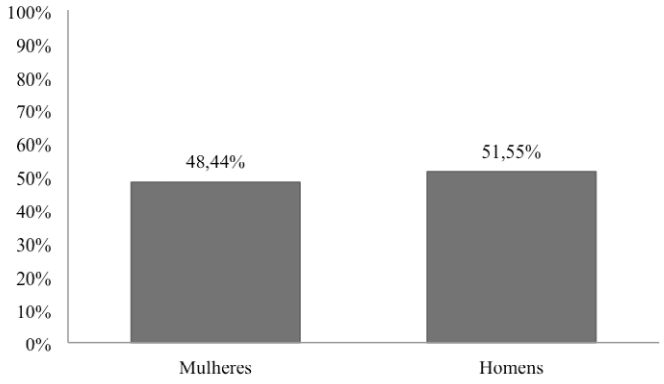


Gráfico 1 – Estratificação da população alagoana por sexo

De acordo com a estratificação por faixa etária, apresentada pela PNAD-2009, a população alagoana, em 15 grupos etários, conforme gráfico 2, apresenta percentual maior de pessoas entre o grupo de 10 a 14 anos, representando 11,1% da população, seguido do grupo de 15 a 19 anos, com um percentual de 10,7%. Já o menor grupo é formado por pessoas com idade entre 65 e 69 anos, representando 2,4% da população alagoana, seguido do grupo de 60 a 64 anos com 3% (Gráfico 2).

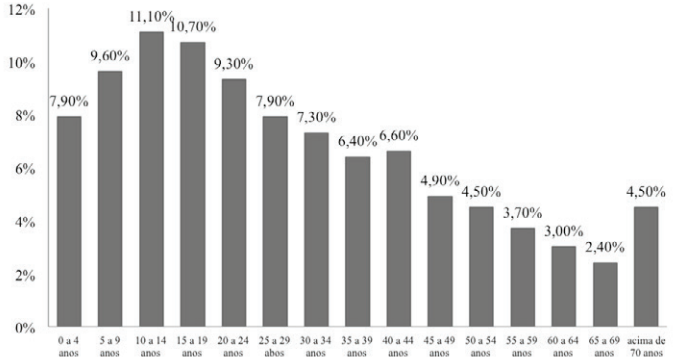


Gráfico 2 – Estratificação da população alagoana por faixa etária.

Quanto aos anos de escolarização, o PNAD – 2009, considerando apenas as pessoas de 10 anos ou mais de idade, aponta que pessoas com 4 a 7 anos de estudos constituem o maior grupo da população alagoana, representando 30,1%; já as pessoas com 15 anos ou mais de estudos representam apenas 4,1% (Gráfico 3).

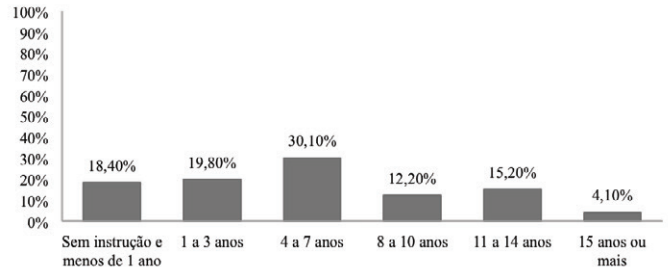


Gráfico 3 – Estratificação da população alagoana por anos de estudos.

Os dados do PNAD – 2009 também apontam que mais de 68,3% da população alagoana com 10 anos ou mais de idade não possuem o ensino fundamental completo, e que 18,4% estão sem instrução ou possuem menos de um ano de estudo. Esses dados sociodemográficos nos embasam para o dimensionamento da amostra sociolinguística.

2. ESTRATIFICAÇÃO DA AMOSTRA

Selecionada a comunidade de fala, o passo seguinte foi a estratificação da amostra a ser coletada. Para tanto, partimos do pressuposto de que o termo

amostra "refere-se ao grupo de indivíduos [...] selecionados para *representar*, no estudo, a população ou o universo do qual fazem parte e que o pesquisador deseja estudar" (GUY; ZILLES, 2007, p. 109, grifo nosso).

> Há, pois, uma pressuposição de que o comportamento lingüístico dos indivíduos, cujo discurso examinamos reflete regularidades ligadas ao fato de que aderem às normas de seus respectivos grupos sociais; é nesse sentido que os resultados do estudo do comportamento de certo número de indivíduos (a amostra) são generalizados para os grupos sociais aos quais eles pertencem (e representam). (GUY; ZILLES, 2007, p. 109).

De acordo com Guy e Zilles (2007, p. 109), algumas perguntas norteiam a constituição de uma amostra, a saber: "Como definir, identificar ou delimitar os grupos sociais que constituem uma comunidade? Quais deles devem ser incluídos na amostra? Como relacionar os indivíduos necessários para ter uma amostra representativa nesse sentido estatístico?".

Embora não haja uma resposta simples e única para essas questões, pois muitas alternativas têm sido adotadas por diferentes pesquisadores para a realização do trabalho de campo sociolinguístico, o que direcionam as respostas a esses questionamentos são os objetivos propostos em cada pesquisa sociolinguística. Os critérios de constituição de uma amostra devem ser coerentes com a pesquisa que se pretende realizar. Para a constituição de nossa amostra, estabelecemos, logo de início, dois parâmetros para a seleção dos informantes: os falantes deveriam ser pessoas nascidas em Alagoas e que não tivessem se afastado do Estado por tempo superior a cinco anos.

Em seguida, estratificamos a amostra em três células sociais: sexo, faixa etária e escolaridade, e as subdividimos nos seguintes fatores: sexo (masculino / feminino), faixa etária (F1 – 15 a 29 anos / F2 – 30 a 44 anos / F3 – mais de 44 anos) e escolaridade (E1 – Ensino Fundamental / E2 – Ensino Médio / E3 – Ensino Superior). Com isso, obtivemos um total de 18 células, conforme Quadro 1.

Masculino	F1	E1	Feminino	F1	E1
Masculino	F1	E2	Feminino	F1	E2
Masculino	F1	E3	Feminino	F1	E3
Masculino	F2	E1	Feminino	F2	E1
Masculino	F2	E2	Feminino	F2	E2
Masculino	F2	E3	Feminino	F2	E3
Masculino	F3	E1	Feminino	F3	E1
Masculino	F3	E2	Feminino	F3	E2
Masculino	F3	E3	Feminino	F3	E3

Quadro 1 – Estratificação da amostra. Fonte: Vitório (2012, p. 68).

A partir da estratificação, delimitamos o número de informantes necessários para obtermos uma amostra representativa da comunidade estudada. Selecionamos quatro informantes por células e obtivemos um total de 72 (4 x 18 = 72) a serem entrevistados, constituindo, assim, uma amostra sincrônica composta por 72 entrevistas.

De acordo com Guy e Zilles (2007), em uma pesquisa de cunho variacionista, o ideal é selecionar quatro ou cinco informantes em cada célula, para evitar, durante o momento da entrevista e constituição do *corpus* da pesquisa, um comportamento linguístico idiossincrático ou enviesado, caso trabalhássemos com um ou dois informantes por células.

> O acréscimo de uma terceira pessoa já nos daria chance de identificar as tendências de uso para aquele grupo, mas ainda assim poderíamos enfrentar dúvidas relacionadas com diferenças (se são por acaso, por idiossincrasia ou por razões de outra ordem) e ter pouca base para fazer qualquer tipo de generalização. Por isso, diz-se que, com 4 ou 5 indivíduos em cada célula, aumentamos substancialmente as chances de identificar *tendências* através da constatação de regularidades no comportamento dessas pessoas, em contraste com o de outras pessoas da amostra. (GUY; ZILLES, 2007, p. 112-113, grifo nosso).

Quanto à seleção desses informantes, seguimos o método aleatório estratificado, levando em consideração não só que eles deveriam ser pessoas nascidas em Alagoas e que não tivessem se afastado do Estado por tempo superior a cinco anos, mas também deveriam corresponder às células sociais definidas na estratificação da amostra.

Nesse método, o pesquisador sociolinguista divide a amostra em células sociais (no nosso estudo, dividimos em sexo, faixa etária e escolaridade) compostas, cada uma delas, por indivíduos que apresentam as mesmas características sociais e em seguida, procede-se, a partir de uma seleção aleatória, ao preenchimento de cada célula.

Também consideramos, conforme Milroy e Milroy (1992), a abordagem "bola de neve", em que um informante, a partir do contato com o pesquisador, indica outro para fazer parte da pesquisa. A adoção dessa abordagem também facilitou o acesso aos informantes, uma vez que mediava a interação entre entrevistador e entrevistado.

3. COLETA DOS DADOS

Delimitada a amostra da pesquisa, o passo seguinte foi a coleta dos dados. Guy e Zilles (2007, p.20) apontam essa atividade lida com as seguintes perguntas:

"Como obtemos os dados? Os dados são válidos para refletir o fenômeno investigado? Os procedimentos para a obtenção dos dados são confiáveis e reproduzíveis? O que pode ser feito para minimizar a parcialidade dos dados?".

Para a obtenção dos dados, elaboramos uma ficha da amostra sociolinguística contendo os dados dos informantes a serem entrevistados: nome, naturalidade, profissão, sexo (masculino / feminino), faixa etária (F1 – 15 a 29 anos / F2 – 30 a 44 anos / F3 – mais de 44 anos) e escolaridade (E1 – Ensino Fundamental / E2 – Ensino Médio / E3 – Ensino Superior). A elaboração da ficha social, além de fornecer informações sobre os informantes, o familiariza o com gravador e mapeia seus possíveis interesses, auxiliando no momento da entrevista.

Também elaboramos um questionário-guia de entrevistas, com os tópicos da conversa:

1. Fale-me da sua profissão/curso.
2. Como é o seu dia de trabalho/estudo? O que você faz?
3. Pretende fazer algum outro curso? Qual? Por quê?
4. Fale-me um pouco da sua cidade e da administração do atual prefeito.
5. Fale-me um pouco do nosso estado e da administração do atual governador.
6. Qual a sua opinião sobre essa violência toda que está havendo aqui?
7. O que você faria para amenizá-la?
8. Você (amigo/parente/conhecido) já sofreu algum tipo de violência? O que aconteceu?
9. Fale-me de um passeio/viagem que você fez e achou interessante.
10. Já passou por alguma situação que pôs sua vida em risco? O que aconteceu?

O questionário-guia de entrevistas tem como principal objetivo homogeneizar os dados para posterior comparação, controlar os tópicos da conversa e provocar narrativas de experiências pessoais, tendo em vista que estudos com narrativas de experiências pessoais têm demonstrado que, ao relatá-las, o informante está tão envolvido com *o que* fala que presta o mínimo de atenção ao *como* fala, o informante "fica envolvido na narrativa a ponto de parecer estar revivendo aquele momento." (LABOV, 2008, p. 119).

Em seguida, entramos em contato com os informantes e realizamos as entrevistas, que foram feitas, na maioria das vezes, em nosso primeiro encontro, devido à disponibilidade e ao interesse dos falantes em participar da pesquisa[2].

2 Como nosso objeto de estudo era a realização dos verbos *ter* e *haver* em construções existenciais na fala alagoana, verificamos, durante o período de coleta de dados, que as entrevistas realizadas, tanto no primeiro contato com os informantes quanto no segundo, não interferiam na realização da variação em estudo.

Nossas entrevistas aconteceram ou nas residências, ou nos locais de trabalho dos informantes, no período de fevereiro a julho de 2010.

Nesse período de coleta de dados, entrevistamos aproximadamente 15 informantes por mês. No entanto, convém ressaltar que, ao considerarmos o método aleatório estratificado na seleção dos informantes, realizamos mais entrevistas no início do que no final da coleta de dados, uma vez que à medida que as células sociais foram sendo preenchidas, aumentavam ainda mais as especificidades dos informantes a serem entrevistados.

Como equipamento de gravação, utilizamos um gravador digital padrão e armazenamos nossas entrevistas no formato de arquivo .wav. Para obtermos um material linguístico em que predominasse a espontaneidade da fala dos entrevistados, nossas interferências geralmente ocorriam para estimular a continuidade da fala. O resultado de todas as gravações, de mais ou menos 15 minutos por falantes, corresponde a aproximadamente 1080 minutos de falas, totalizando quase 18 horas de entrevistas[3].

> [...] nosso objetivo é observar o modo como as pessoas usam a língua quando não estão sendo observadas. Todos os nossos métodos envolvem uma aproximação a esse objetivo: quando fazemos uma abordagem a partir de duas direções diferentes e obtemos o mesmo resultado, podemos ter certeza de que conseguimos vencer o paradoxo do observador no sentido de que a estrutura existe independentemente do analista. (LABOV, 2008, p. 83).

Realizadas todas as entrevistas, obtivemos uma amostra da comunidade de fala alagoana composta de 72 informantes, oriundos das regiões destacadas no mapa da Figura 1.

3 Entendemos que entrevistas sociolinguísticas de base laboviana duram em média de 50 a 80 minutos, desconsiderando, na maioria das pesquisas, os 15 minutos iniciais de coleta de dados, devido ao estranhamento em relação ao contexto de entrevista. No entanto, como nosso foco nessa coleta era a variação dos verbos *ter* e *haver* em construções existenciais, a gravação de mais ou menos 15 minutos por informantes nos forneceu um *corpus* estatisticamente relevante para a análise dos dados.

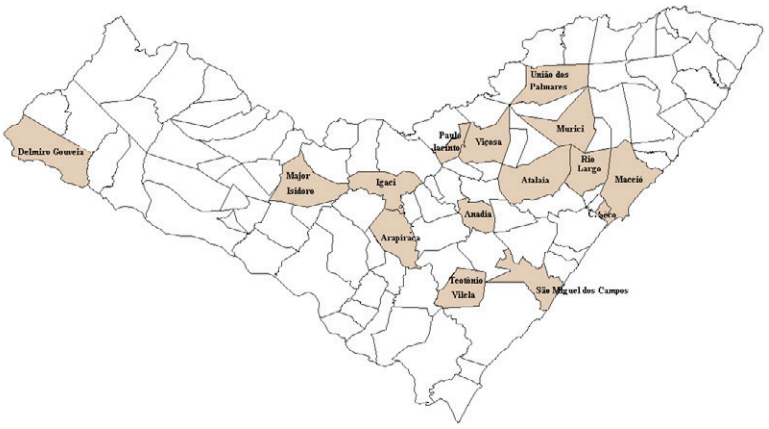


Figura 1 – Distribuição geográfica dos informantes. Fonte: Vitório (2012)

Embora não consideremos, na estratificação da amostra, a origem geográfica dos informantes, entrevistamos, no momento da coleta de dados, falantes que residiam em várias regiões do estado, para que pudéssemos ter uma visão geral da língua usada em Alagoas.

4. TRANSCRIÇÃO DOS DADOS

Após a realização das entrevistas, o passo seguinte foi transcrevê-las. Na prática, as transcrições dos dados ocorreram simultaneamente às entrevistas, ou seja, ao fim de cada, procurávamos logo iniciar o trabalho de transcrição dos dados. Procedemos dessa forma com o objetivo de capturar de forma mais fidedigna possível os fatos relatados.

Para tanto, seguimos o Protocolo de Transcrição do Projeto A Língua Usada em Alagoas (LUAL)[4], segundo o qual, todas as entrevistas gravadas tiveram transcrição ortográfica, ou seja, procuramos seguir a ortografia oficial, mas registrando o máximo de questões características da fala coletada, conforme podemos observar no excerto transcrito a seguir.

4 As convenções de transcrição foram adotadas da adaptação do modelo da Equipe do Groupe Aixois de Recherches en Sociolinguistique (GARS), dirigido por Blanche-Benveniste, para o português, realizadas pela Profª Drª Denilda Moura, para o projeto A Língua Usada em Alagoas – LUAL. Tanto as convenções de transcrição quanto o Protocolo de Transcrição do Projeto LUAL estão disponíveis no banco de dados do Programa de Extensão em Ensino de Línguas (PRELIN), no Programa de Pós-Graduação em Letras e Linguística da Universidade Federal de Alagoas.

> L27– cara eu acho que a principal causa é:: é a falta de imprego e oportunidade pru povo porque não tem educação de qualidade – você num tem educação de qualidade você num tem incentivo – porque as pessoas xxx porque como lá nu colégio que eu tava dando aula você via – os alunos gritava – ah vô istudá pra que? vô mi formá e num vô tê imprego – vô mi formá e num máximo que vô consegui é vendê no shopping – aí eu vô tá feliz cum isso – eu acho que isso é uma das grandes causas – num tem um um alvo – a educação num presta – você num tem um objetivo pra xxx

As transcrições dos dados foram feitas com o auxílio do programa computacional Express Scribe, que pode ser obtido gratuitamente na internet e objetiva auxiliar o pesquisador na tarefa de transcrição do registro de áudio. A escolha desse programa foi motivada por possibilitar o uso de teclas de atalhos para executar diversas funções, permite recortar pedaços registrados da fala, facilitando o processo de edição, e visualiza as informações trabalhadas em uma única janela, conforme Figura 2.

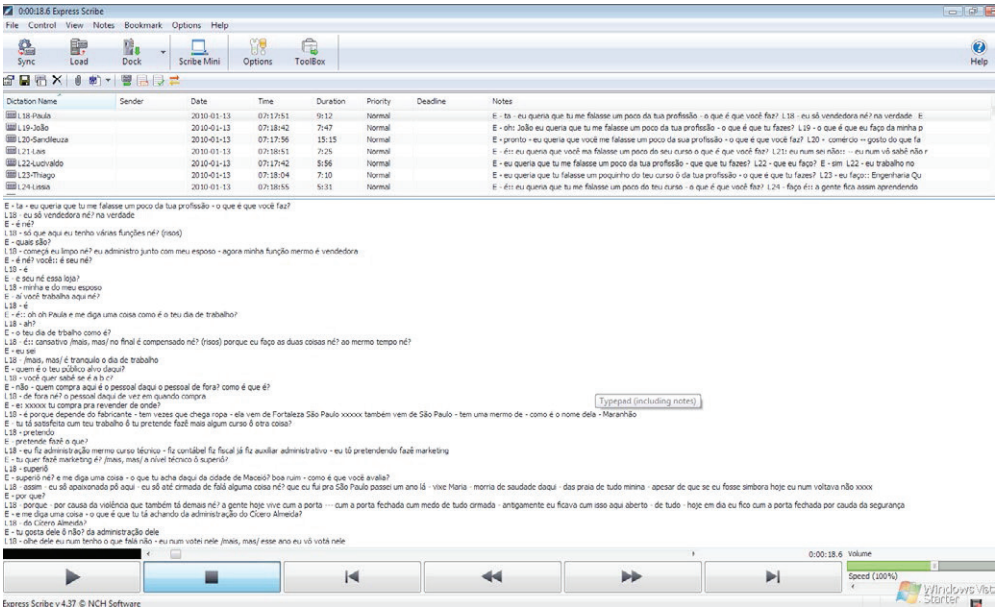


Figura 2 – Janela de trabalho do Express Scribe

Realizadas as transcrições das entrevistas, fizemos uma leitura de revisão para checar se os dados transcritos estavam de acordo com as falas coletadas.

> O pesquisador deve ter em mente que, embora a transcrição seja o resultado do que é percebido e, portanto, também é consequência do cruzamento das opções teóricas do

> pesquisador, da interpretação atribuída aos dados e dos objetivos da pesquisa, não existe uma transcrição irretocável, por isso é conveniente que a transcrição realizada possa ser 'avaliada' por outra(s) pessoa(s) e, quando necessário, submetida a revisões. (DE PAULA, 2011, p. 37).

Atualmente, está sendo realizada mais uma nova checagem das transcrições realizadas e, quando terminada, a amostra sincrônica da fala alagoana, coletada e transcrita, será disponibilizada à comunidade acadêmica científica como mais uma fonte para estudos descritivos de variedades do português brasileiro falado.

CONCLUSÕES

Adotando os pressupostos teórico-metodológicos da Teoria da Variação e Mudança, apresentamos os passos metodológicos seguidos para a coleta de uma amostra sincrônica da língua falada em Alagoas. Para tanto, focalizamos nossa discussão na delimitação da comunidade estudada, na estratificação da amostra a ser coletada e nas tarefas de coleta e transcrição dos dados. Vale ressaltar que, mesmo seguindo os passos da Sociolinguística Quantitativa, os procedimentos aqui adotados tinham como objetivo a pesquisa sobre a variação dos verbos "ter" e "haver" em construções existenciais na fala alagoana (VITORIO, 2012). Dessa forma, consideramos que "cada pesquisa impõe determinados limites, obrigando o pesquisador à contorná-los como uma forma de obter dados pertinentes, confiáveis e representativos do universo pesquisado" (DE PAULA, 2011, p. 39).

REFERÊNCIAS

CAMPOY, J; ALMEIDA, M. *Metodología de la investigación sociolingüística*. Granada: Editorial Comares, 2005.

BRASIL. Ministério do Planejamento, Orçamento e Gestão. Instituto Brasileiro de Geografia e Estatística. *Censo 2010*. Brasília, 2010. Disponível em: <www.Ibge.gov.br>. Acesso em: set. 2010.

_____. Ministério do Planejamento, Orçamento e Gestão. Instituto Brasileiro de Geografia e Estatística. *Pesquisa Nacional por Amostras de Domicílios* – PNAD/2009. Brasília: 2009. Disponível em: <www. Ibge.gov. br>. Acesso em: set. 2010.

DE PAULA, A. O trabalho de campo sociolinguístico. In: COSTA, J.; SANTOS, R.; VITÓRIO, E. (org.). *Variação e mudança linguística no estado de Alagoas*. Maceió: EDUFAL, 2011, p. 29-41.

GUY, G.; ZILLES, A. *Sociolinguística quantitativa*: instrumental de análise. São Paulo: Parábola Editorial, 2007.

LABOV, W. *Padrões sociolinguísticos*. Trad. M. Bagno, M. M. P. Scherre, C. R. Cardoso. São Paulo: Parábola Editorial, 2008. Título original: Sociolinguistic Patterns, 1972.

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

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

VITORIO, E. *Ter/haver existenciais na fala alagoana: variação estável ou mudança em progresso?*. Alagoas, 2012. Tese (Doutorado em Linguística). Faculdade de Letras, Universidade Federal de Alagoas.

CAPÍTULO 5

PROCEDIMENTOS METODOLÓGICOS PARA UMA INVESTIGAÇÃO SOCIOLINGUÍSTICA COM A LÍNGUA BRASILEIRA DE SINAIS

Liliane Correia Toscano de Brito Dizeu

INTRODUÇÃO

A língua de sinais, usada pelas comunidades de surdos no Brasil, é basicamente produzida com as mãos, embora movimentos do corpo e da face desempenhem diferentes funções. Por ser uma língua de modalidade gesto-visual, a Língua Brasileira de Sinais (Libras) faz uso de movimentos gestuais e expressões faciais, que são percebidos pela visão (PEREIRA, 2000). É reconhecida legalmente como língua, como sistema linguístico legítimo, e não como um problema do surdo ou patologia da linguagem (QUADROS; KARNOPP, 2004).

Quanto à estrutura, tanto as línguas de sinais quanto as orais apresentam as mesmas propriedades abstratas da linguagem, mas se opõem fortemente em suas formas na superfície. Os estudos de Stokoe (1960) mostraram que os sinais não são somente imagens, mas símbolos abstratos, possuindo uma complexa estrutura interior. Foi o pioneiro na investigação para buscar a estrutura, analisar os sinais e dissecá-los e a pesquisar suas partes constituintes.

Stokoe primeiramente comprovou que cada sinal é constituído por pelo menos três partes independentes (em analogia com os fonemas da fala): o ponto

de articulação, a configuração das mãos e o movimento, e que cada uma dessas partes apresentam um número limitado de combinações.

Os articuladores primários das línguas de sinais são as mãos, que se movimentam no espaço em frente ao corpo e articulam sinais em determinados pontos. Um sinal pode ser articulado com uma ou duas mãos (QUADROS; KARNOPP, 2004).

O ponto de articulação dos sinais é o espaço em frente ou em uma região do próprio corpo. Os sinais articulados são de dois tipos: os que se articulam no espaço neutro diante do corpo e os que se aproximam de uma determinada região, tais como: cabeça, mão, cintura e os ombros (FERREIRA BRITO, 1995). Já os movimentos podem envolver uma vasta rede de formas e direções, desde os internos da mão, os do pulso e os direcionais no espaço (QUADROS; KARNOPP, 2004).

Durante muito tempo, as línguas de sinais foram consideradas apenas como gestos, incapazes de expressar conceitos abstratos. Só foram reconhecidas como línguas quando surgiu um sistema de notação que pudesse representar a estrutura de seus sinais. As pesquisas sobre as línguas de sinais são muito recentes, se comparadas as línguas orais, que já apresentam uma longa tradição. Além disso, a maioria delas ainda não estão totalmente descritas em seus níveis fonológico, morfológico e sintático e carecem de maior investigação. Com relação à Libras, as pesquisas linguísticas ainda são escassas, e há a necessidade de mais trabalhos na área para ampliar a sua descrição.

Em pesquisa realizada com a Língua de Sinais da Nova Zelândia, Mckee e Mckee (2006) observaram a ocorrência de variação sociolinguística no nível lexical, possivelmente decorrentes do modelo educacional adotado: grupos de sinalizadores mais velhos apresentaram diferenças no uso da língua de sinais, se comparados a um grupo de jovens sinalizadores. O grupo de surdos mais velho teve uma experiência educacional a partir de um modelo voltado para a oralidade, enquanto o grupo mais jovem vivenciou um período educacional em que a língua de sinais foi introduzida nas escolas e a comunidade surda da Nova Zelândia passava por um processo natural de mudança de língua.

A partir deste estudo, surgiu o interesse de verificar como ocorre a variação linguística na Libras, tendo em vista o processo educacional desenvolvido na cidade de Maceió, em Alagoas, decorrente da mudança política e social sofrida por esta língua. Para tanto, apresentamos uma metodologia de coleta de dados desenvolvida para os dados de Libras.

1. PROCEDIMENTO DE COLETA PARA ANÁLISE DA VARIAÇÃO EM LIBRAS

A partir do estudo de Mckee e Mckee (2006) e com base na realidade local, foram selecionados 50 itens lexicais de cinco categorias semânticas, como: alimento, animal, cor, meio de transporte e vestuário. As figuras foram dispostas em tamanho A4, com imagens reais e coloridas. O léxico selecionado foi adaptado de acordo com a realidade dos sujeitos, utilizando figuras comuns na região, por não haver um modelo de pesquisa nacional nem com a língua portuguesa e nem com a Libras.

As figuras correspondiam a: *alimento* (carne, queijo, macarrão, maçã, morango, caju, abacaxi, pão, chocolate e pipoca); *animal* (macaco, borboleta, boi, cavalo, cachorro, tartaruga, sapo, gato, leão e galinha); *meio de transporte* (bicicleta, moto, carro, caminhão, navio, avião, carroça, trem, helicóptero e ônibus); *vestuário* (bolsa, relógio, vestido, gravata, sapato, calcinha, camisa, calça, sandália e chapéu) e *cores* (marrom, azul claro, azul escuro, laranja, amarelo, cinza, vermelho, preto, verde e rosa).

A metodologia usada na constituição do *corpus* foi a utilizada pelos trabalhos de sociolinguística a partir da perspectiva variacionista (LABOV, 2008). A identificação de fenômenos variáveis pressupõe que, para os membros de uma mesma comunidade de fala, existam pelo menos duas possibilidades de representação superficial para uma determinada categoria linguística. A escolha entre as formas não se dá de maneira aleatória ou livre, mas relacionada às variáveis linguísticas e extralinguísticas. As variações são as modificações que surgem em um ou mais parâmetros da Libras, quando apenas um ou dois são modificados temos uma variação fonológica, quando a modificação ocorre nos três podemos dizer que há uma variação lexical. Nas Figuras 1 e 2, verificamos a variação fonológica no item HELICÓPTERO.

Figura 1 – HELICÓPTERO – var. 1

Figura 2 – HELICÓPTERO – var. 2

Na Figura 1, a mão está aberta e a mão direita está em formato da letra D, enquanto na Figura 2 a mão esquerda permaneceu com a mesma configuração e a mão direita modificou, assumindo a configuração da letra L do alfabeto manual. Os parâmetros ponto de articulação e movimento não sofrem modificações.

A coleta de dados envolveu a participação de 18 sujeitos, sendo estratificados quanto ao sexo e idade, conforme Quadro 1.

Sexo	
Masculino	9 informantes
Feminino	9 informantes
Faixa etária	
15 a 23 anos	6 informantes
24 a 32 anos	6 informantes
33 a 41 anos	6 informantes

Quadro 1 – Extratificação da amostra.

Inicialmente, foi prevista uma amostra com 24 sujeitos, contudo não foi encontrado um sujeito do sexo feminino com faixa etária entre 33 a 41 anos. Restringiu-se, portanto, a amosta a 18 sujeitos surdos, com o seguinte perfil: perda neurossensorial profunda, ausência de comorbidades (como a dificuldade cognitiva), usuários da Libras e do português escrito, com escolaridade entre o ensino fundamental e superior completo, participantes da comunidade surda, residentes na cidade de Maceió.

A coleta de dados teve início após aprovação do Comitê de Ética e Pesquisa da Universidade de Ciências da Saúde de Alagoas, segundo o protocolo de número 786, e durando quatro meses, em apenas um encontro com cada participante. A coleta foi realizada em escolas, associação e centros especializados para surdos, previamente marcada, conforme a disponibilidade dos informantes, e contou com uma aluna do Curso de Fonoaudiologia da UNCSAL ouvinte, que interpretou e explicou em Libras o texto do Termo de Consentimento Livre e Esclarecido para os sujeitos, bem como o procedimento. No encontro foi preenchida a ficha social para verificar os critérios de inclusão e posterior classificação a partir dos aspectos extralinguísticos (sexo e idade) e, em seguida, as figuras foram mostradas, uma por uma, com a pesquisadora solicitando aos participantes que realizassem o sinal correspondente a cada ilustração.

Todos os encontros foram documentados com uma câmera filmadora, o que possibilitou, posteriormente, a análise dos dados. A filmagem de cada informante realizando os 50 sinais durou aproximadamente cinco minutos. Os informantes foram filmados da região da cintura até o topo da cabeça e das extremidades de um braço ao outro, o que permitiu a visualização total da realização dos sinais.

Apesar dos sinais de cores terem sido coletados, foram excluídos da amostra, pois vários informantes apresentaram dificuldades para realizar a distinção entre eles. A amostra constituída é composta por 900 sinais realizados pelos sujeitos, com cerca de cinco minutos de gravação para cada informante, totalizando 90 minutos.

Foram consideradas como variações as modificações que surgiram nos sinais dos informantes em relação aos aspectos fonológicos em um ou mais parâmetros, visto que qualquer modificação do movimento, da configuração das mãos ou do ponto de articulação pode alterar o significado do sinal apresentado. Inicialmente, para definir o parâmetro de análise, foi utilizado o Dicionário Enciclopédico Trilíngue da Língua de Sinais Brasileira, volumes I e II (CAPOVILLA; RAPHAEL, 2001), por ser um material de referência construído a partir de dados coletados nas principais capitais brasileiras. Como a língua recebe várias interferências regionais, um objeto pode apresentar mais de um sinal, dependendo da localização geográfica em que é utilizado.

Durante a análise dos dados observamos que, além dos sinais variarem em relação ao dicionário, apresentavam variação também entre si. Dessa maneira, optamos por verificar a variação apenas entre os pares, em detrimento do uso do dicionário.

2. ANÁLISE DA VARIAÇÃO EM LIBRAS

Para analisar os dados, os sinais foram descritos e, em seguida, verificada a presença ou não de variação, que quando presente, foi correlacionada aos fatores sociais (idade e sexo).

Análises descritivas indicam que 27,5% dos sinais foram realizados com modificações no ponto de articulação, ou seja, dos 40 sinais pesquisados, 11 apresentaram variação nesse parâmetro. As categorias semânticas e os respectivos itens lexicais que sofreram variação neste parâmetro foram: alimentos (queijo, caju, morango e pão); animais (macaco e galinha); transporte (carroça) e vestuário (bolsa, sapato, camisa e calça). Vejamos exemplos dos efeitos dos fatores sociais na variação dos sinais.

2.1. Variável sexo

Para o item SAPATO, a variante mais frequente (Figura 3) apresentou maior número de ocorrência no sexo feminino e na segunda e terceira faixa etária.

Figura 3 – SAPATO var. 1	Figura 4 – SAPATO var. 2	Figura 5 – SAPATO var. 3
		

A variante da Figura 3 apresenta uma realização mais cuidadosa do sinal em detrimento das outras, o que corrobora com a constatação de Mckee, Mckee e Major (2004) de que as sinalizantes femininas realizam os sinais com maior destreza, sendo a questão articulatória um fator relacionado ao sexo/gênero do sujeito. A segunda variante foi realizada apenas pelo sinalizantes masculinos.

Nas Figuras 3, 4 e 5 observa-se a diferença sutil na execução do sinal, que não compromete o acesso lexical das variantes, identificando o seu significado pelo interlocutor.

2.2. Acréscimo lexical

Quadros e Karnopp (2004) afirmam que os sinais com pontos de articulação que apresentam mais de um elemento (região principal) em suas representações são considerados complexos. E, segundo Lucas, Bayley e Valli (2003), a variação fonológica afeta as partes básicas dos sinais, essas unidades podem ser alteradas, adicionadas, removidas ou rearranjadas. Observa-se em alguns sinais a adição, o que ocorreu em quatro dos onze sinais, gerando variação no ponto de articulação.

A primeira variante para CARROÇA (Figura 6) evidencia que houve uma adição no sinal, a fim de delimitar o léxico. A segunda variante foi representada pelo sinal da Figura 7.

Figura 6 – CARROÇA var.1	Figura 7 – CARROÇA var.2
	 

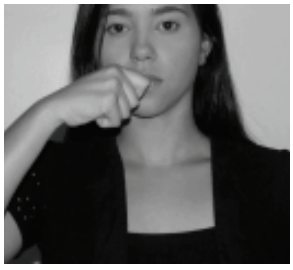
Na segunda variante, sujeitos do sexo feminino apresentaram maior frequência de realizações, apesar de ser considerada uma variante simples. No entanto, o acréscimo não implica uma elaboração mais refinada e sim uma necessidade de melhor explicar o sinal. Fuentes e Tolchinsky (2004, apud MCKEE; MCKEE; MAJOR, 2004) afirmam que um maior número de variantes pode ocorrer dependendo do perfil social dos informantes, exemplificando que, se os sujeitos forem professores, pode ocorrer um número maior de variantes acadêmicas. Geralmente, os professores procuram elaborar melhor o sinal, o que pode ser o caso, visto que alguns dos sujeitos que participaram da coleta são professores de Libras e um deles apresentou acréscimos em três dos quatros sinais realizados. Esses profissionais tendem a querer explicar melhor o sinal, adicionando mais características e realizando as variantes com acréscimo.

A tentativa de melhor caracterização também pode ocorrer devido ao fato de os sujeitos surdos estarem realizando o sinal para um interlocutor que é ouvinte, como uma forma de facilitar a compreensão, o que pode não ocorrer se o diálogo for entre surdos. Mckee, Mckee e Major (2004) explicam que os informantes podem não realizar o sinal de forma espontânea, dependendo da forma da coleta de dados.

2.3. Variação etária

Sandler (1989, apud Quadros; Karnopp 2004) afirma que o ponto de articulação principal, produzido no espaço neutro, é o local de articulação de um sinal que estará associado a dois subespaços (subníveis). Esses subníveis identificam e distinguem os itens lexicais, assim como os traços distintivos nas línguas orais.

O item lexical PÃO caracteriza-se pela variante frequente ocorrer na região principal do espaço neutro, e as demais serem realizadas em uma região do corpo. As variantes do sinal apresentam subníveis de uma mesma região principal.

Figura 8 – PÃO var. 1	Figura 9 – PÃO var. 2	Figura 10 – PÃO var. 3
		

A variante mais frequente (Figura 8) foi realizada no espaço neutro em frente à boca (50%), sendo encontrada em todos os informantes da terceira faixa etária. É possível que essa variante desapareça brevemente, já que é utilizada por poucos sujeitos das faixas etárias mais jovens. Tais dados sugerem evidente mudança para a segunda variante (33,3%), realizada pela maioria dos sujeitos da primeira faixa etária.

A lei nº 6.060 de 15 de setembro de 1998, do estado de Alagoas, reconheceu e implantou a Libras na rede pública de ensino para a comunidade surda. Em 24 de abril de 2002 foi decretada a lei nº 10.436, que regulamentou a Libras como a forma de comunicação e expressão da comunidade de pessoas surdas do Brasil. Tais dispositivos legais sugerem que apenas os sujeitos da primeira faixa etária possam ter aprendido Libras nas escolas, com intérpretes, e esse fato contribui para minimizar a variação entre esses sujeitos.

Mckee, Mckee e Major (2004) sugerem que o aprendizado institucional da língua sinalizada vem a refletir um padrão de mudança em direção à substituição de formas fonologicamente mais elaboradas, nesse caso apresentando o ponto de articulação em uma região principal e não apenas no espaço neutro, como pode ser observado nas Figuras 8, 9 e 10.

O parâmetro ponto de articulação apresenta interferência na variação encontrada na Libras, embora seja ainda um parâmetro pouco abordado na literatura. Atualmente, o curso de licenciatura e bacharelado em Letras/Libras tem o objetivo de formar professores e tradutores/intérpretes de Libras/Português e o desenvolvimento da metodologia de coleta aqui apresentada é importante para a descrição da variação linguística entre a comunidade surda do Brasil.

CONCLUSÕES

As variações do parâmetro ponto de articulação da Libras ocorreram por diferenças nas regiões principais, nos subníveis e por acréscimo de outro ponto de articulação. Há variações que comprometem e as que mantêm a compreensão do sinal, e o fator idade apresentou maior influência na variação do que o sexo, sinalizando as variantes inovadoras e as que tendem a desaparecer com o tempo.

A metodologia desenvolvida possibilitou a coleta de dados de Libras e pode ser replicada, o que pode contribuir para o avanço nos estudos sociolinguísticos brasileiros.

REFERÊNCIAS

ALAGOAS. Lei n. 6.060, de 15 de Setembro de 1998. Dispõe sobre o reconhecimento e a implantação da Linguagem Brasileira de Sinais-LIBRAS como língua oficial na rede pública de ensino para surdos, e adota providências correlatas. *Gabinete civil do estado de Alagoas*: Palácio Marechal Floriano, Maceió, 110º da república, 1998.

BRASIL. Lei n. 10.436, de 24 de Abril de 2002. Dispõe sobre a Língua Brasileira de Sinais - Libras e dá outras providências. *Planalto do governo central:* Congresso Nacional, Brasília, 2002; 181º da Independência e 114º da República.

CAPOVILLA, F. C.; RAPHAEL, W. D. *Dicionário enciclopédico ilustrado trilíngüe da Língua de Sinais Brasileira*. São Paulo: EDUSP, 2001. 1 v.

_____. *Dicionário enciclopédico ilustrado trilíngüe da Língua de Sinais Brasileira*. São Paulo: EDUSP, 2001. 2 v.

LABOV, W. *Padrões sociolinguísticos*. Trad. M. Bagno, M. M. P. Scherre, C. R. Cardoso. São Paulo: Parábola Editorial, 2008. Título original: Sociolinguistic Patterns, 1972.

LUCAS, C.; BAYLEY, R.; VALLI, C. What's your sign for pizza?: an introduction to variation. In: *American sign language*. Washington: G. U. Press, 2003.

MCKEE, D.; MCKEE, R. Investigating sociolinguistic variation in New Zeland sign language. In: QUADROS, R. M. *Congresso Internacional de Aspectos Teóricos das Pesquisas nas Línguas de Sinais*. Florianópolis: Lagoa Editora, 2006, p.127 -128; 314-328.

PEREIRA, M. C. C. A Língua de Sinais na Educação de Surdos. In: LACERDA, C. B. F.; NAKAMURA, H.; LIMA, M. C. (Org.). *Fonoaudiologia: surdez e abordagem bilíngue*. São Paulo: Plexus Editora Ltda. 2000, p. 13-20.

QUADROS, R. M. de.; KARNOPP, L. B. *Língua de Sinais Brasileira:* estudos linguísticos. Porto Alegre: Artmed, 2004.

STOKOE, W. Sign language structure: An outline of the visual communication systems of the American deaf. *Studies in linguistic*, n.8, 1960.

CAPÍTULO 6

O BANCO DE DADOS FALA-NATAL: UMA AGENDA DE TRABALHO

Maria Alice Tavares
Marco Antonio Martins

INTRODUÇÃO

Estudos feitos na perspectiva da Sociolinguística Variacionista vêm fomentando a ampliação do conhecimento sobre o português brasileiro, desde a década de 1970, através da descrição e da análise de fenômenos variáveis nos âmbitos fonológico, morfológico, sintático, semântico e discursivo. Contudo, há estados da federação em que tais pesquisas são ainda incipientes ou mesmo inexistentes. É o caso do Rio Grande do Norte (RN), que não conta com um banco de dados de fala com as características necessárias para a pesquisa sociolinguística. Para suprir essa lacuna, propusemo-nos a organizar um *corpus* de fala denominado Banco de Dados da Fala do Rio Grande do Norte (FALA-RN), que contará com amostras representativas de diferentes comunidades de fala norte-rio-grandenses.

O marco inicial da organização do Banco de Dados FALA-RN é a constituição do Banco de Dados FALA-Natal, que congrega entrevistas sociolinguísticas feitas com membros da comunidade de fala e, na medida do possível, de diferentes comunidades de prática, do município de Natal, que é a capital e o maior centro urbano do estado potiguar. Posteriormente, serão coletadas entrevistas sociolinguísticas em comunidades de fala representativas do interior do RN.

Apresentamos, neste texto, um balanço geral da constituição do banco de dados FALA-Natal para o qual as entrevistas estão em fase final de coleta.

Apresentamos ainda, as dificuldades práticas com as quais temos nos deparado na constituição do *corpus*, assim como as soluções encontradas.

1. O BANCO DE DADOS FALA-NATAL

Os informantes do Banco de Dados FALA-Natal são socialmente estratificados de modo similar aos informantes de bancos de dados já existentes no país, a exemplo do Programa de Estudos sobre o Uso da Língua (PEUL) e do projeto Variação Linguística na Região Sul do Brasil (VARSUL). Além disso, buscamos considerar aspectos relacionados à identificação de comunidades de prática em que estejam engajados os informantes.

Inicialmente, o Banco de Dados FALA-Natal será composto por 48 entrevistas sociolinguísticas com cerca de 60 minutos de duração. Essas entrevistas serão distribuídas, em termos de estratificação social, quanto ao sexo (24 informantes de sexo feminino e 24 informantes de sexo masculino); idade (12 informantes de 8 a 12 anos, 12 informantes de 15 a 21 anos, 12 informantes de 25 a 50 anos e 12 informantes de mais de 50 anos) e nível de escolaridade (12 informantes com ensino fundamental I completo, 12 informantes com ensino fundamental II completo e 12 informantes com ensino médio completo, além de 12 informantes cursando o ensino fundamental I – os indivíduos de 8 a 12 anos). Serão entrevistados informantes de diferentes bairros das quatro zonas de Natal.

Essa constituição inicial do Banco de Dados busca similaridade com os de dados sociolinguísticos já existentes no Brasil, com o objetivo de facilitar a execução de estudos sociolinguísticos comparativos. Entre as sugestões para futuras investigações e desdobramentos, frequentemente apontadas em pesquisas concluídas na perspectiva da Sociolinguística Variacionista, destaca-se a possibilidade de realização de análises comparativas dos resultados obtidos para o fenômeno estudado com os de pesquisas efetuadas em outras regiões do Brasil, que tenham o mesmo ou semelhante objeto de estudo.

No entanto, essas análises sociolinguísticas comparativas de grande extensão são pouco realizadas, talvez por conta das dificuldades que sua execução implica. Conforme Guy (1999), muito é perdido ao se deixar de empreender comparações entre resultados obtidos para fenômenos variáveis dentro de uma mesma língua ou mesmo interlinguísticas, pois um dos objetivos centrais da Sociolinguística Variacionista é o estabelecimento de princípios gerais e, na medida do possível universais, que estariam subjacentes à variação e à mudança, e seriam válidos para todas ou grande parte das comunidades de fala.

Segundo Tagliamonte (2003, p. 729), a "comparação sempre esteve na raiz da sociolinguística", permitindo a construção de generalizações através do

cotejamento de amostras de dados em tempo real e em tempo aparente. Todavia, análises comparativas interdialetais apenas recentemente vêm recebendo um destaque crescente no cenário mundial, o que tem levado à proposição de generalizações e mesmo de princípios universais de variação e mudança. Como afirma Chambers (2004, p. 128), "à medida que a sociolinguística se torna menos restrita a eventos locais, se torna comparativa e, à medida que o aspecto comparativo ganha peso, generalizações interlinguísticas não apenas se tornam possíveis, mas inevitáveis".

Assim, para fazer avançar ainda mais a sociolinguística no Brasil, é de fundamental importância a realização de estudos comparativos que visem buscar semelhanças e diferenças quanto ao comportamento de uma mesma variável linguística, em diferentes dialetos brasileiros. Quais os ganhos que adviriam desse tipo de estudo?

Entre muitas vantagens, a principal parece ser a possibilidade de observar se as restrições linguísticas e sociais à variação e à mudança para um dado fenômeno são as mesmas em todas as regiões do Brasil e, se não, em que diferem, aventando explicações que abranjam resultados provindos de diversas comunidades de fala. A partir de tais observações e explicações, podemos chegar a estabelecer generalizações, base dos princípios gerais tão procurados pela Sociolinguística Variacionista para a construção de sua teoria. Contudo, não é apenas sobre o que é comum às comunidades de fala que recai o interesse de um estudo comparativo: a comparação pode auxiliar na descoberta de especificidades e de idiossincrasias em comunidades particulares, revelando o jogo local *versus* universal típico da língua.

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

2. SOBRE AS DIFICULDADES ENCONTRADAS E O ENCAMINHAMENTO DE SOLUÇÕES

Na constituição do banco de dados FALA-Natal, que ainda está em desenvolvimento (em fase final de realização das entrevistas), temos nos defrontado com uma série de questões para as quais temos buscado soluções. Entre essas questões, apontamos as seguintes:

- **Representatividade da amostra:** De acordo com o Censo de 2010 do IBGE, a capital norte-rio-grandense tem 803.739 habitantes. Se considerarmos uma amostragem na condição metodológica ideal, aplicando o corte de 0,5% da população, o banco de dados FALA-Natal deveria contar com 4.019 entrevistas/falantes.

  Em condições reais, o desenvolvimento de um banco de dados com esse número de entrevistas demandaria anos de realização. Com o significativo crescimento da população, se adotássemos essa condição para o desenvolvimento do banco, quando a última entrevista fosse realizada, a comunidade, com certeza, já não seria a mesma. Além disso, a quantidade de informantes também depende de financiamento e de quanto tempo se dispõe para a organização do banco de dados, fatores que, em geral, impedem a coleta de um grande número de entrevistas.

  De qualquer forma, um número menor de entrevistas pode ser representativo de tendências gerais da comunidade. Segundo Sankoff (1988, apud TAGLIAMONTE, 2006, p. 23), é necessário "não que a amostra seja uma versão em miniatura da população, mas apenas que tenhamos a possibilidade de fazer inferências sobre a população com base na amostra". Cada banco de dados deve ter um mínimo de representatividade com base em idade, sexo, classe social e/ou nível de educação, o que assegura que a diversidade linguística da comunidade de fala esteja representada na amostra.

  Lembramos que a maior coleta de entrevistas sociolinguísticas já feita foi dirigida por Shuy et al. (1968), tendo sido gravadas 702 entrevistas em Detroit, nos Estados Unidos. No entanto, as análises mais detalhadas desse *corpus* utilizaram apenas 48 dessas entrevistas, com os informantes distribuídos simetricamente em quatro classes sociais, em um total de 12 informantes por classe (cf. WOLFRAM, 1969; LABOV, 2008; CHAMBERS, 1995; TAGLIAMONTE, 2006).

  No caso do Brasil, os bancos de dados costumam ter de 2 a 3 informantes por célula social, o que tende a ser suficiente para a obtenção dos padrões gerais de variação de uma comunidade de fala no que diz respeito aos diversos fenômenos variáveis. Quanto ao Banco de Dados FALA-Natal, caso algumas características de uso linguístico variável chamem, por alguma razão, a atenção no conjunto das 48 entrevistas sociolinguísticas iniciais, outras entrevistas poderão ser realizadas – com os mesmos ou outros informantes – no sentido de possibilitar uma análise mais refinada desses usos.

- **Dificuldade de localização de informantes com certos traços socioeconômicos:** A esse respeito, nossa maior dificuldade atualmente está na localização

de indivíduos com mais de 50 anos, nascidos em Natal e com pais natalenses. Boa parte dos indivíduos que temos contatado são oriundos do interior do Rio Grande do Norte. Para tentar resolver esse problema, estamos utilizando a metodologia de "bola de neve", pedindo a cada informante com mais de 50 anos que nos indique amigos e/ou conhecidos com as mesmas características sociais;

- **Necessidade de maior diferenciação de faixas etárias para testar hipóteses relativas à mudança linguística:** No Banco de Dados FALA-Natal, estamos considerando quatro faixas etárias, prevendo 12 informantes de 8 a 12 anos, 12 informantes de 15 a 21 anos, 12 informantes de 25 a 50 anos e 12 informantes de mais de 50 anos. A proposta de termos informantes com menos de quinze anos é motivada pela possibilidade de descoberta, para casos de mudança linguística, de padrões caracterizados por um pico de uso na fala dos adolescentes.

  Como vários estudos sociolinguísticos vêm constatando a existência do uso intenso de formas inovadoras por indivíduos em torno de dezesseis a vinte anos de idade, Labov (2001) propôs a existência de um pico de uso de formas inovadoras no período final da adolescência. Segundo Labov, é esperado que ocorra, nos processos de mudança, após o pico de uso da forma inovadora, uma retração de seu aparecimento: ela é incorporada, ainda com índices de grande frequência, à gramática dos falantes do grupo em que teve seu uso fortemente acelerado, mas passa a recorrer menos, em comparação com a fase de pico de uso. Desse modo, a mudança adquire matizes não tão radicais, e sim, uma maior gradualidade: passa a haver uma distribuição linear crescente ou decrescente entre faixas etárias adultas, agora ocupadas pelos mesmos indivíduos que levaram a forma inovadora a seu ápice. A forma poderá vir a derrotar as demais concorrentes com o passar do tempo, mas com uma menor velocidade do que a prevista, considerando-se somente seu estágio de pico de uso.

  Para que seja possível a verificação, em cada fenômeno variável para o qual tenhamos indícios de mudança, dessa possibilidade de ocorrência da forma inovadora na fala adolescente, é preciso que levemos em conta faixas etárias menores. No caso do Banco de Dados FALA-Natal, estamos contando com uma faixa etária de 8 a 12 anos, adicionando assim, os pré-adolescentes ao banco de dados.

- **Validade da comparação entre análises realizadas em dados extraídos de entrevistas sociolinguísticas feitas recentemente** (o que será o caso do Banco de Dados FALA-Natal), **e dados extraídos de entrevistas sociolinguísticas**

feitas há dez ou vinte anos: Essa é outra questão que nos preocupa, afinal, na comparação de resultados provenientes de bancos de dados constituídos em diferentes períodos de tempo, sempre será necessário ter em mente que cada comunidade de fala comparada pode estar representando uma etapa diferente de variação e mudança linguística. Por exemplo, se comparássemos dados extraídos do VARSUL referentes à comunidade de fala de Florianópolis (cujo banco foi organizado ao longo da última década do século XX) com os bancos constituídos recentemente em outras comunidades de fala, seria preciso levar em conta, na análise dos resultados, que os bancos mais novos podem estar representando um período mais recente de evolução para certos fenômenos variáveis em processo de mudança em comparação com o banco de Florianópolis.

- **Observação de aspectos relacionados à questão da análise estilística pelo viés da terceira onda**: Em relação a esse último tópico, nosso objetivo inicial foi a coleta de entrevistas sociolinguísticas em uma comunidade de fala ampla para que seja possível a realização de mapeamentos de tendências gerais de variação e mudança em relação a essa comunidade.

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

  Também foram propostos, nas entrevistas, tópicos que estimulassem o informante a discorrer sobre as diferentes comunidades de prática das quais faz parte. Elaboramos um roteiro para as entrevistas com sugestões de perguntas que o entrevistador poderia fazer ao entrevistado. Entre elas, estão questões do tipo: (i) Com quem você passa o tempo, além das pessoas da sua família? O que vocês fazem juntos? Que tipo de lazer vocês têm?; (ii) Você participa de algum trabalho voluntário? Como é?; (iii) Você participa de algum grupo de jovens? O que vocês fazem juntos?; (iv) Você participa de algum grupo da igreja? Como é?; (v) Você frequenta algum clube? Qual? Como é?; (vi) Algo interessante já aconteceu no clube/grupo de jovens/

grupo da igreja quando você estava? O que aconteceu? e (vii) Descreva o que você faz em um dia, desde que acorda até ir dormir.

Esse maior conhecimento sobre as comunidades de prática em que se engaja cada informante, que será obtido através das fichas sociais e das próprias entrevistas poderá ser considerado na análise dos fenômenos variáveis. Tanto as informações extraídas das fichas sociais e das entrevistas, não apenas fornecerão subsídios para uma análise mais aprofundada de cada informante, no que tange às características sociais e de prática, como também trarão indícios a respeito de quais comunidades de prática – entre as inúmeras em que se integra cada indivíduo – são mais interessantes para a realização de estudos nos moldes da terceira onda, o que pode, naturalmente, ensejar a coleta de novas entrevistas com grupos de indivíduos pertencentes a tais comunidades.

Ou seja, para a organização do Banco de Dados FALA-Natal, estamos conscientes da necessidade de contemplar não apenas pesquisas sociolinguísticas afiliadas à abordagem variacionista de Labov (um retrato amplo de comunidades de fala definidas geograficamente), mas também à abordagem etnográfica alinhada ao Milroy (um retrato local, etnográfico, de comunidades de fala definidas geograficamente) e à abordagem da identidade social alinhada à Eckert (um retrato do(s) indivíduo(s) integrante(s) de comunidades de prática, pelo viés do estilo como elemento central de constituição da *persona*).

3. UM BALANÇO DA CONSTITUIÇÃO DO CORPUS E UMA AGENDA DE TRABALHO

Com a organização do Banco de Dados FALA-RN, do qual a composição do Banco de Dados FALA-Natal representará a primeira etapa, será possível a descrição de dialetos do português brasileiro falados em um estado nordestino no qual, sob a perspectiva da variação e da mudança linguística, pouco foi feito. Para preencher essa lacuna, o Banco de Dados FALA-RN fomentará o desenvolvimento de projetos voltados para a pesquisa, o ensino e a extensão, tanto nos cursos de pós-graduação quanto nos de graduação, bem como oferecerá aos interessados em geral uma fonte de dados linguísticos contemporâneos. Estamos em fase de conclusão das entrevistas que constituirão o banco FALA-Natal. As entrevistas, em formato digital, serão tratadas e ficarão à disposição da comunidade acadêmica para pesquisas.

REFERÊNCIAS

CHAMBERS, J. Dynamic typology and vernacular universals. In: KORTMANN, B. (Ed.). *Dialectology meets typology: dialect grammar from a cross-linguistic perspective*. Berlin: Mouton de Gruyter, 2004. p. 127-145.

GUY, G. R. Notas do curso Sociolinguística Comparativa, ministrado de 22/02 a 05/03, na UFSC, por ocasião do XIV Instituto Linguístico da ABRALIN, 1999.

LABOV, W. *Padrões sociolinguísticos*. Trad. M. Bagno, M. M. P. Scherre, C. R. Cardoso. São Paulo: Parábola Editorial, 2008. Título original: Sociolinguistic Patterns, 1972.

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

TAGLIAMONTE, S. A. Comparative sociolinguistics. In: CHAMBERS, J. K.; TRUDGILL, P.; _____. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press, 2006.

SHILLING-ESTES, N. (Eds.). *The handbook of language variation and change*. Cambridge: Blackwell, 2003. p. 729-763.

SHUY, R.; WOLFRAM, W.; RILEY, W. *Field techniques in an urban language study. Washington, DC*: Center for Applied Linguistics, 1968.

WOLFRAM, W. *A sociolinguistic description of Detroit Negro speech*. Washington, DC: Center for Applied Linguistics, 1969.

7 CAPÍTULO

REDES SOCIAIS, IDENTIDADE E VARIAÇÃO LINGUÍSTICA

Elisa Battisti

INTRODUÇÃO

Estudos de comunidades em pequena escala, como as análises de rede social, "são capazes de fornecer informação mais detalhadas sobre o uso que os falantes fazem da variabilidade linguística", em especial no que se refere "às partes menos formais do repertório linguístico" (MILROY, 1980, p.21). A concentração das relações sociais nas redes, em um dado território, concorre para o desenvolvimento do sentimento de pertença da identidade local, construída através da relativa homogeneidade de comportamento – no vestir, no falar, no divertir-se, no alimentar-se, nos valores praticados, entre outros – como assume o estudo da variação na linha das práticas sociais (ECKERT, 2000). Este trabalho, sobre "Redes sociais, identidade e variação linguística", traz alguns fundamentos teóricos e uma análise variacionista na rede social, no intuito de explicitar os procedimentos metodológicos necessários a esse tipo de investigação.

Labov (2010), ao discutir os fatores sociais que, em seu conjunto, dirigem o processo de mudança linguística e moldam a história da divergência dialetal, afirma que rede social e comunidades de prática são duas das forças motrizes da variação e mudança. Unidades sociais menores do que a classe, essas forças dão relevo ao indivíduo no processo de mudança. O autor afirma que redes de maior complexidade e densidade preservam falares contra os efeitos do nivelamento dialetal, e que os líderes da mudança são os membros da rede com o

maior número de contatos dentro e fora dela. Sobre as comunidades de prática, esclarece que a variação é usada para evocar diferentes identidades e, na negociação dos indivíduos por *status*, as formas linguísticas adquirem valor social, o que pode incrementar ou fazer regredir a mudança. Reconhecida a pertinência de investigar redes e práticas sociais no estudo da variação linguística, resta aos sociolinguistas o desafio de dar conta dessas forças, associando as medidas quantitativas da análise de regra variável (LABOV, 1972) a outras técnicas de investigação, assentadas em claros fundamentos teóricos.

1. IDENTIDADE

Apesar de as identidades serem experimentadas, vivenciadas pelos sujeitos e, nas investigações, serem consideradas pelo exame das práticas sociais individuais, elas são em parte construtos sociais. Como explica Bonnewitz (2003), viver em sociedade implica socialização, isto é, aprendizagem de normas, valores e crenças de coletividades que pautam suas práticas, suas ações e comportamentos. Na perspectiva de Bourdieu (1977), socializar-se é realizar essa aprendizagem interiorizando normas, valores e crenças, como um sistema de disposições estruturantes duradouras, isto é, como princípios geradores e organizadores de práticas e representações que regulam tacitamente a ação cotidiana (*ethos*) e as posturas corporais (*hexis*). Em outras palavras, socializar-se é adquirir o *habitus*. Com essa aquisição, os sujeitos tornam-se seres sociais e suas identidades individuais vão sendo definidas:

> O *habitus* está na base daquilo que, no sentido corrente, define a personalidade de um indivíduo. Nós mesmos temos a impressão de termos nascido com essas disposições, com esse tipo de sensibilidade, com essa maneira de agir e reagir, com esses modos e com esse estilo. Na verdade, gostar mais de cerveja do que de vinho, de filmes de ação do que de filmes políticos, votar mais na direita do que na esquerda são produtos do *habitus*. Do mesmo modo, andar com o tronco erguido ou curvado, ser desajeitado ou ter facilidade nas relações interpessoais são manifestações da *hexis* corporal. [...] considerar determinado indivíduo como pequeno, mesquinho, ou, pelo contrário, generoso, brilhante, depende do *ethos*. (BONNEWITZ, 2003, p.78)

É importante esclarecer que o *habitus* é um sistema de disposições, não de determinações estruturantes. Pela socialização, adquirimos tendências a agir, a nos comportarmos e pensarmos de dadas formas, mas não ficamos fadados a isso. Vale dizer, então, que, embora tácitos, o *habitus* e nossas identidades estão sujeitos a mudanças e ajustes derivados de novas formas de participação em comunidades ao longo da

vida. É assim que, segundo Wenger (1998), a construção de identidade consiste em negociar os significados de nossa experiência de pertença a diferentes grupos sociais. Nossas identidades são fruto de nossa filiação social, das posições que ocupamos nos grupos de que fazemos parte, esses estruturados em relação aos campos ou classes sociais distintas. Nas palavras de Bonnewitz (2003, p.91), os comportamentos e ações sociais que nos parecem mais naturais são "apenas o produto de múltiplas aquisições sociais: a personalidade individual é apenas uma variante de uma personalidade social constituída na e pela filiação a uma classe social".

Na mesma linha, Wenger (1998) esclarece que nossas identidades não são apreendidas somente pelas nossas práticas sociais, são também relativas à nossa posição e à posição de nossas comunidades na estrutura social mais ampla. Em termos analíticos, então, isso torna desnecessário escolher entre a comunidade ou a pessoa como unidade de análise. "O foco deve estar no processo de sua constituição mútua", defende a autora (WENGER, 1998, p.146).

Nessa perspectiva, a das práticas sociais (WENGER, 1998), identidade é então, (i) vivida: não é uma categoria, traço de personalidade, papel ou rótulo, é uma experiência que envolve participação e reificação; (ii) negociada: é um permanente vir a ser, não é definida apenas em um período específico da vida; (iii) social: é fruto da pertença a grupos; (iv) processo de aprendizagem: é uma trajetória no tempo que incorpora o presente, o passado e o futuro; (v) nexo: combina múltiplas formas de participação; (vi) local-global: não se constrói apenas pelas práticas imediatas ou se regula somente pelas estruturas sociais mais amplas, é uma interface de ambas.

No estudo de variação linguística como prática social de Eckert (2000), essa foi a perspectiva de identidade seguida. As identidades dos membros das comunidades de prática investigadas numa escola em Detroit (EUA), Jocks e Burnouts, foram analisadas relativamente à pertença ao grupo, às formas de participação dos membros e às práticas por eles realizadas. Os grupos, por sua vez, foram percebidos em relação às classes sociais dos pais dos alunos e seu estatuto no cenário escolar. Para tanto, a autora empregou a técnica etnográfica da observação participante e fez análise de rede social, além da análise quantitativa dos dados. Essa é a ideia que fica do trabalho de Eckert (2000), integrar técnicas de análise que aproximem o pesquisador da comunidade investigada, sem abrir mão dos procedimentos (quantitativos), tradicionalmente empregados na pesquisa variacionista.

Entretanto, apesar da validade dos resultados obtidos por Eckert (2000), as técnicas de pesquisa, principalmente a etnografia com observação participante, podem demandar muito tempo do investigador e assim, serem custosas do ponto de vista pessoal e financeiro, o que talvez inviabilize projetos de pesquisa de orçamento e cronograma limitados. A sugestão que fica é a de, na medida do possível, manter a essência do estudo de variação linguística em que práticas sociais e

identidade têm papel. Isso pode significar reduzir o tempo de observação, mas não deixar de fazê-la. Sugere-se usar o ingresso na comunidade, necessário para a realização das entrevistas sociolinguísticas, também como momento de observação participante; seguir roteiros de entrevista que, ao abordar temas do cotidiano, levem os informantes a discorrer sobre a comunidade, sua participação em grupos, suas práticas sociais diárias; dependendo dos resultados da análise quantitativa para os grupos de fatores sociais, identificar estratos sociais que mereçam investigação posterior, como comunidade de prática. Por exemplo, sendo o fator gênero feminino o condicionador de um processo variável e as informantes integrantes do clube de mães, realiza-se observação participante por algum tempo no clube.

Para a análise de rede, o ideal é identificar as conexões entre os membros dos grupos, também com observação participante, como fez Eckert (2000). Se isso mostrar-se inviável, sugere-se que a rede se forme quando, da realização das entrevistas sociolinguísticas, um informante indica outro com o perfil requisitado pelo pesquisador. À medida que a rede se forma, o quadro dos sujeitos já entrevistados vai sendo apresentado aos informantes subsequentes, que dizem se conhecem ou não os anteriores, e qual é sua forma de relacionamento. Esses procedimentos têm respaldo teórico na própria noção de rede social, como veremos a seguir.

2. REDE SOCIAL

Milroy e Milroy (1992) afirmam que, ao engajarem-se em grupos, as pessoas criam uma estrutura significativa para a resolução dos problemas que surgem em seu cotidiano. Como bem observa Eckert (2000, p.34), "os problemas diários mudam, assim como as pessoas". Embora possam relacionar-se localmente com mais intensidade, conhecendo quase todos os membros de uma comunidade e esses, conhecendo-se também, os indivíduos movimentam-se, engajam-se em diferentes empreendimentos e em variadas comunidades, nas quais processos simbólicos e relações identitárias diversas têm lugar. Há ligação entre redes e práticas sociais na variação e mudança linguística.

No prefácio à segunda edição do trabalho pioneiro de Milroy (1980) sobre redes sociais e variação linguística, "Language and Social Networks" (Língua e Redes Sociais), Peter Trudgill observa que a proposta articula um estudo laboviano em dialetologia social a um estudo socioantropológico e psicossocial da língua, na esteira de Gumperz.

Estudos de rede social não são método exclusivo à análise linguística. Nas Ciências Sociais, as redes têm sido analisadas desde a década de 1970. Em Castells (1999), as redes são representações da morfologia de organização social da sociedade contemporânea, especialmente das redes informacionais. Trata-se

de uma categoria de pesquisa mais flexível, menos comprometida com as generalizações universais, mais próxima à dimensão do cotidiano.

Consideradas como teias de laços que se estendem, potencialmente, a toda a sociedade, as redes apresentam diferenças em sua configuração estrutural em duas dimensões, a da *densidade* e a da *plexidade* (do inglês *density* e *plexity*, respectivamente). Conforme Evans (2004), a densidade (estrutura da rede) refere-se aos contatos dos indivíduos: quanto maior o número de pessoas em rede que se conhecem, maior sua densidade. Já plexidade (conteúdo da rede) à multiplicidade de conexões dos membros. Por exemplo, pode ter membros que sejam vizinhos (rede uniplexa), ou também colegas de escola (rede multiplexa).

As redes sociais são ancoradas nos indivíduos. Li Wei (1996) afirma que, por essa razão, geralmente interessam as análises das redes, cujos laços estabelecem-se entre pessoas que interagem diretamente, o que limita a um número entre 20 e 50 o total de participantes da rede analisada. Ainda, distinguem-se laços fortes dos fracos: opõem-se, respectivamente, laços "que conectam amigos e parentes, àqueles que conectam conhecidos" (MILROY, 2002, p.550).

Conforme Evans (2004), as redes sociais podem ser vistas tanto como um sistema de relações pessoais com efeitos sobre os indivíduos ou como relações usadas pelas pessoas para atingir seus objetivos. A primeira visão é a mais frequentemente adotada por sociolinguistas, entre eles Milroy (1980).

Milroy (1980) faz uma análise do vernáculo, ou modalidade não padrão da língua[1]. A autora estudou três bairros de classe trabalhadora (*working-class neighborhoods*) de Belfast em seus padrões variáveis de realização vocálica, predominantemente. A quantificação da variação, correlacionada à rede social dos informantes, revelou que o emprego majoritário de alternantes vernaculares reflete os padrões de interação social entre as comunidades em redes densas, multiplexas. Esses padrões não poderiam ser explicados por gênero, idade e classe social, dada a homogeneidade das comunidades em rede. A autora complementa: "diferentemente das classes sociais mais abstratas, esses grupos sempre têm uma forte base territorial" (MILROY, 1980, p.14). Os bairros de classe trabalhadora investigados por Milroy são habitados por pessoas que, em função de limitadas condições socioeconômicas, não apresentam grande mobilidade territorial. Interagem socialmente no próprio bairro o que contribui para desenvolverem um forte sentimento de pertença a ele, como se fossem os proprietários daquela área da cidade. A esse sentimento de pertença ao bairro, e o valor social (positivo) a ele atribuído Milroy denomina *localismo*. Na interação local, os contatos de uma

1 Denominação corrente nos estudos filiados à Sociolinguística Variacionista laboviana para as modalidades de fala menos formais, em que se verificam relativos desvios à norma gramatical ou às variedades mais prestigiadas socialmente.

pessoa conhecerão uns aos outros, integrando uma rede social densa, e quase sempre multiplexa, o que sustenta e explica a emergência das variantes vernaculares. Já nas grandes cidades, afirma Milroy (1980), as redes sociais dos indivíduos de alto *status* social (por suas condições econômicas) são mais "abertas", no sentido de que eles se movem para além das fronteiras de seu território, não conhecendo uns aos outros. As relações entre esses indivíduos são regidas por uma supranorma, representada linguisticamente pela modalidade-padrão da língua, menos permeável às mudanças vernaculares.

Milroy (1980, p.16-17) afirma que o conhecimento dos padrões e conflitos (identitários) das comunidades "são extremamente úteis a um investigador da língua", uma vez que lhe "permite dar conta das diferenças sistemáticas no uso da linguagem entre indivíduos e entre subgrupos". Vem daí a ideia de que, em termos analíticos, a medida da densidade e da plexidade da rede venha acompanhada de algum procedimento que capte conflitos e padrões identitários dos grupos pesquisados.

Nas pesquisas que vínhamos realizando, além de associar a análise quantitativa ao estudo de rede social, procedimentos etnográficos, como momentos de observação participante, são empregados para identificar práticas sociais sistematicamente relacionadas às questões culturais. O conteúdo das entrevistas sociolinguísticas é analisado em categorias como: trabalho, religião, transporte, lazer, entre outras, que emergem da fala dos informantes, contribuem para interpretar diferenças estruturais na rede social e diferenças linguísticas entre seus membros. É o que ilustraremos com a retomada da análise de Battisti, Dornelles Filho e Lucas (2009), a seguir.

3. A PALATALIZAÇÃO DAS OCLUSIVAS ALVEOLARES EM ANTÔNIO PRADO (RS)

A palatalização das oclusivas alveolares no português brasileiro (*tia~tʃia, dia~dʒia, mate~matʃ, cidade~cidadʒ*), regra variável desencadeada pelo segmento vocálico anterior alto, derivado ou não, seguinte à consoante, é entendida como processo que se difunde a partir dos centros urbanos (NOLL, 2008). No Rio Grande do Sul, esse é o padrão que se tem verificado (BATTISTI; GUZZO, 2009): alta frequência de aplicação em Porto Alegre (em torno de 90%), frequência moderada em municípios do interior.

Em Antônio Prado, pequeno município gaúcho fundado por imigrantes italianos no final do século XIX, Battisti, Dornelles Filho, Lucas e Bovo (2007a, 2007b, 2008) verificaram frequência de 30% de palatalização. Levantados 26.600 contextos de

palatalização de 48 entrevistas sociolinguísticas do BDSer[2], de informantes dos dois gêneros, quatro grupos etários (15 a 29, 30 a 49, 50 a 69, 70 ou mais anos de idade), habitantes das zonas urbana e rural de Antônio Prado, os pesquisadores constataram que a regra é condicionada, em termos linguísticos, por vogal fonológica ou não derivada /i/ e consoante-alvo da regra desvozeada /t/; em termos sociais, por jovens e habitantes de zona urbana. Considerando-se os resultados da variável idade, poder-se-ia pensar que a mudança estivesse em progresso na comunidade. Porém, comparados os grupos etários, obtém-se um padrão característico às mudanças que se completam. Surge daí a questão: por que a palatalização, apesar de condicionada pelos jovens, mostra sinais de estabilizar-se na comunidade em índices modestos?

3.1. Palatalização em Antônio Prado: a variável idade

A palatalização é moderada em Antônio Prado, aplica-se com uma frequência de 30%. O condicionamento social é bastante forte, desempenhado por jovens e habitantes de zona urbana. À primeira vista, levando-se em conta apenas o comportamento dos grupos etários controlados na variável idade, seria possível afirmar que o processo estaria progredindo na comunidade: a frequência de palatalização aumenta à medida que a idade decresce, como se vê na Figura 1:

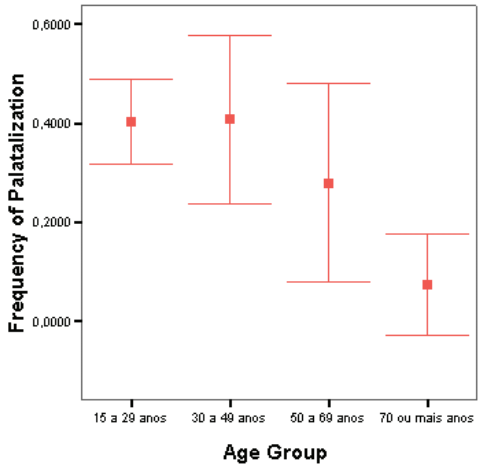


Figura 1 – Frequência de palatalização por grupo etário3 (BATTISTI, DORNELLES FILHO, LUCAS; BOVO, 2008)

2 Banco de Dados de Fala da Serra Gaúcha, do Centro de Ciências Humanas e Comunicação/ Mestrado em Letras e Cultura Regional da Universidade de Caxias do Sul.

3 As barras indicam o intervalo de confiança (IC 95%) para a frequência populacional.

Porém, comparados os grupos etários, forma-se uma linha ascendente (aumento da frequência de palatalização do grupo etário mais idoso para os dois imediatamente precedentes) com um platô no final, decorrente da manutenção da taxa de palatalização do segundo grupo etário para o primeiro, o dos mais jovens. Essa curva em S reflete, nas palavras de Labov (1994, p.65), "... a observação geral de que as mudanças sonoras iniciam-se em uma taxa baixa, progridem rapidamente no seu decorrer e abrandam-se em seus estágios finais."

O que esse padrão representa para a palatalização em Antônio Prado? Ele não permite afirmar que a mudança esteja em progresso na comunidade. O platô se forma em frequências médias de aplicação. Assim, Battisti, Dornelles Filho, Lucas e Bovo (2007a,b, 2008) têm sustentado que, na comunidade, a palatalização tende a estabilizar-se em índices moderados. Como explicar a estabilização? Com a análise de rede, busca-se investigar os inovadores (palatalizadores), que são os indivíduos dos grupos etários mais jovens: posição que ocupam, se periférica ou central, e a qualidade dos laços que ligam os inovadores aos demais membros da rede. O foco baseia-se em Milroy e Milroy (1985), que tratam a difusão da inovação como um aspecto relacionado ao clássico problema da implementação da mudança linguística[4]. A análise da posição e dos laços dos inovadores na rede parte do pressuposto, igualmente inspirado naqueles autores, de que a inovação, adotada inicialmente por membros periféricos (e de laços mais fracos), difunde-se na rede quando os membros centrais (e com laços mais fortes), a adotam. Nossa hipótese é a de que os jovens são membros periféricos na rede, com um maior número de laços, sendo estes mais fracos, unindo-os aos outros integrantes, razão pela qual, apesar de introduzirem a regra e condicionarem a palatalização, o processo não se difunde, estabiliza-se na comunidade em índices modestos. Os idosos são membros centrais na rede, com maior número de laços, sendo estes mais fortes, que reforçam o falar local, sem palatalização. É o que se testa na análise.

3.2. A rede social e a estabilização da palatalização variável em Antônio Prado

Em Battisti, Dornelles Filho, Lucas e Bovo (2007a, 2008), a análise da rede social foi realizada concomitantemente a uma análise de regra variável nos

4 O problema da implementação é uma das cinco áreas de investigação sobre mudança linguística propostas por Weinreich, Labov e Herzog no texto fundador de 1968, *Empirical foundations for a theory of language change* (Fundamentos empíricos para uma teoria da mudança linguística). A questão em torno de que o problema da implementação gira é: por que mudanças num traço estrutural ocorrem numa língua particular num dado período de tempo, mas não em outras línguas com o mesmo traço, ou na mesma língua em outros períodos de tempo?

moldes labovianos. Isso quer dizer que os 48 membros da rede social analisada foram também os 48 informantes, cujas entrevistas foram levantadas os 26.600 contextos de palatalização considerados na análise de regra variável.

A rede social dos informantes foi formada a partir da realização das próprias entrevistas sociolinguísticas, quando um informante indicava outro com as características sociais de interesse[5]. Quando não foi possível obter essa indicação, os pesquisadores, através de seus contatos na comunidade, entrevistaram pradenses com o perfil exigido, a eles perguntando, subsequentemente, se conheciam os demais informantes já entrevistados e que espécie de relacionamento mantinham com cada um.

A rede social foi modelada por um grafo (BOAVENTURA NETTO, 1996; FRUCHTERMAN e REINGOLD, 1991; MATHEWS, 1992; GERHARDT, CORSO e LEMKE, 2005), em que cada informante é um vértice e cada relação de interação é uma aresta. Efetuou-se um estudo do problema de posicionamento dos vértices de modo que o desenho resultante tivesse bons atributos estéticos e de visibilidade computacionalmente elaborados, usando-se uma adaptação do algoritmo de Fruchterman e Reingold (1991) e o método de minimização do gradiente. Cruzaram-se com análise de correlação a frequência de aplicação individual das realizações variáveis, com a aplicação média dos contatos do informante na rede e com características sociais que se mostraram relevantes na análise de regra variável.

Os membros da rede se conhecem e relacionam-se, mas em graus diversos de intimidade. Assim, a rede social dos informantes de Antônio Prado foi analisada em ambas as dimensões, a da densidade e a da plexidade.

Na análise da plexidade, levaram-se em conta graus de relacionamento interindividual ou intimidade/frequência dos contatos. A hipótese seguida foi a de que laços mais íntimos entre os indivíduos implicariam um maior grau de interação pela fala, e isso potencializaria a influência do comportamento linguístico de um indivíduo sobre o do outro. Na reflexão que se empreende aqui, conforme Milroy e Milroy (1985), exercita-se a ideia de que laços mais íntimos sejam laços fortes, menos suscetíveis a mudanças, mais reforçadores do falar de um grupo.

Os graus de intimidade de relacionamento foram inspirados em Blake e Josey (2003), mas adaptados às características da localidade de Antônio Prado, de acordo com as práticas sociais/culturais informadas nas próprias

5 O BDSer selecionou os informantes em cada município conforme os critérios: gênero (masculino, feminino), idade (18 a 30 anos; 31 a 50 anos; 51 a 70 anos; 71 ou mais anos), escolaridade (0 a 4 anos; 5 a 8 anos; 9 a 11 anos; 12 ou mais anos) e local de residência (zona urbana e zona rural).

entrevistas, como também com dados das fichas sociais dos informantes. Os laços familiares e de colegas de trabalho são os relacionamentos mais importantes em Antônio Prado, pela sua intimidade e frequência de interação. No entanto, nem todo o laço desse tipo é igual. No ambiente de trabalho, por exemplo, distinguem-se os que supõem intimidade e interação diária, que não implicam tal frequência e modo de interação. O mesmo se aplica aos laços de amizade, de vizinhança, de colaboração em associações e aos estabelecidos entre parentes, quando, conforme os depoimentos dos próprios entrevistados, distinguem-se parentes próximos de parentes distantes. O Quadro 1 traz os graus considerados na análise: 1, 2 e 3, do mais aos menos íntimos/frequentes, de acordo com padrões locais:

1. Primeiro grau
1A – Marido/mulher
1B – Pais/filhos
1C – Colega de trabalho com interação
2. Segundo grau
2A – Tios/sobrinhos/primos/cunhados
2B – Amigos íntimos
2C – Vizinho íntimo
2D – Colega de associação com interação
3. Terceiro grau
3A – Amigo não-íntimo
3B – Vizinho não-íntimo
3C – Colega de trabalho sem interação
3D – Colega de associação sem interação
3E – Tios/sobrinhos/primos/cunhados sem interação

Quadro 1 – Graus de relacionamento em rede em Antônio Prado (BATTISTI, DORNELLES FILHO, LUCAS; BOVO, 2007a).

A rede social com todos os informantes de Antônio Prado está na Figura 2. Cada um é representado por um retângulo com sua própria frequência de palatalização. Linhas contínuas ligam informantes que possuem um relacionamento de primeiro grau, linhas tracejadas ligam os que têm um relacionamento de segundo grau e linhas pontilhadas ligam os que têm um relacionamento de terceiro grau.

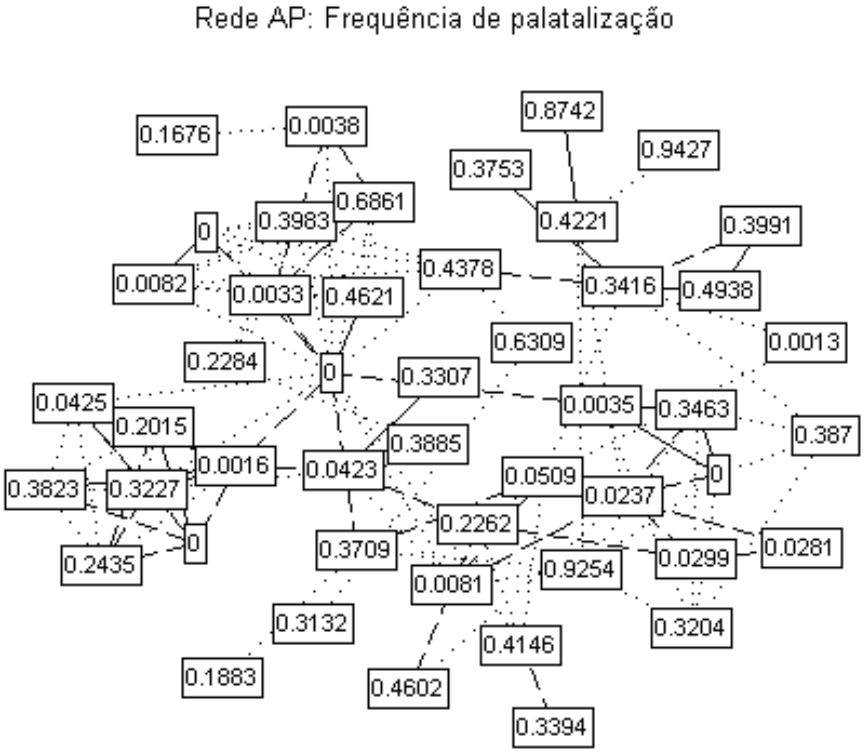


Figura 2 – Rede social dos informantes de Antônio Prado com frequência individual de palatalização (BATTISTI, DORNELLES FILHO, LUCAS; BOVO, 2008).

O conhecimento mútuo é predominante na rede, razão pela qual é considerada densa. Nela integram-se informantes das zonas urbana e rural, como se vê na Figura 3 (por U lê-se zona urbana e por R, zona rural):

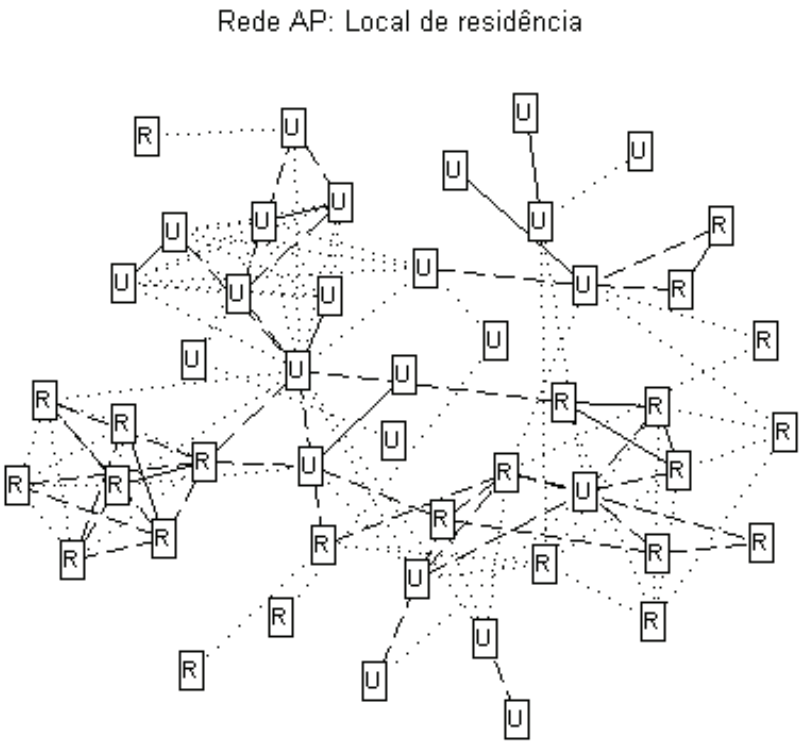


Figura 3 – Rede social dos informantes de Antônio Prado com local de residência.

Embora a rede apresente agrupamentos (ou *clusters*, em inglês) predominantemente urbanos ou rurais, neles há sempre algum membro ligado à outra área.

Quanto à plexidade, constata-se que a quantidade de laços de grau 1 e 2 (mais íntimos ou fortes) é proporcionalmente maior na zona rural do que na zona urbana, embora a diferença não seja estatisticamente significativa. Os laços ligeiramente mais fortes na zona rural fortaleceriam o vernáculo local, o que poderia explicar o caráter desfavorecedor da área à aplicação da regra de palatalização.

E que posição ocupam os inovadores (jovens) na rede, independentemente da zona que habitam? São eles periféricos? Veja-se a Figura 4. Nela, os números de 1 a 4 representam os grupos etários controlados: 1 é o grupo que reúne informantes de 15 a 29 anos; 2, de 30 a 49 anos; 3, de 50 a 69 anos; 4, de 70 ou mais anos. Informantes que se situam no núcleo de agrupamentos são centrais, possuem um maior número de contatos; opostamente, informantes com um menor número de contatos são periféricos.

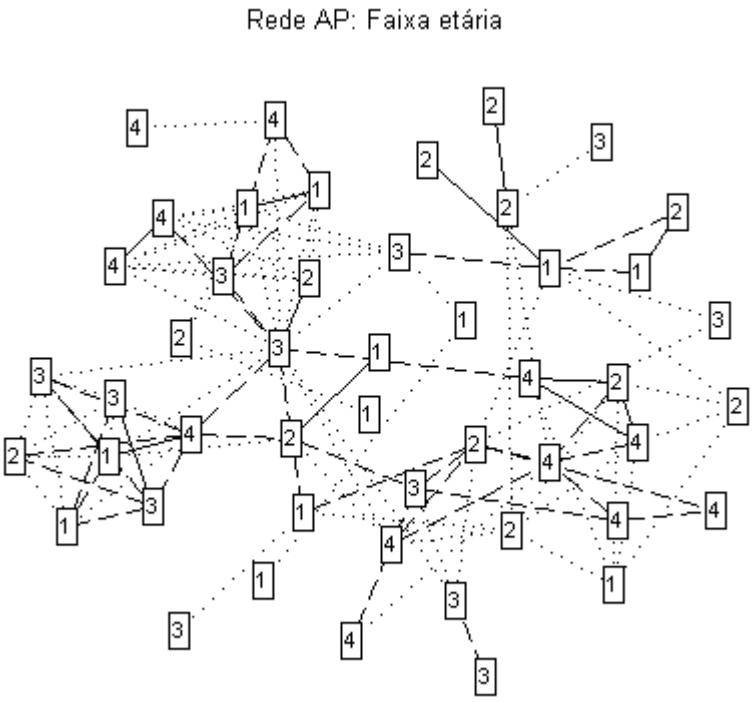


Figura 4 – Rede social dos informantes de Antônio Prado com grupos etários.

Considerando-se os informantes do grupo etário 1, não se pode afirmar que sejam periféricos, tampouco centrais. O número de contatos desses informantes com os demais membros da rede varia de 1 a 9, sendo essa medida similar aos dos informantes dos outros grupos etários. Há, no entanto, um padrão peculiar entre os grupos 1 e 2, que nos permite construir algumas generalizações. Observe-se a Tabela 1. Ela traz a contagem do número de contatos dos informantes por grupo etário e grau de relacionamento.

		Graus de relacionamento	
		Graus 1 e 2	**Grau 3**
Grupos etários	**Grupos 1 e 2**	42 (42%)	65 (52%)
	Grupos 3 e 4	58 (58%)	61 (48%)
	TOTAL	100 (100%)	126 (100%)

Tabela 1 – Número de contatos na rede por grupo etário e grau de relacionamento.

Os grupos etários 1 e 2, juntos, possuem menos contatos de grau 1 e 2 (fortes), mais contatos de grau 3 (fracos), contrariamente aos grupos etários 3 e 4. A diferença não é grande e estatisticamente não significativa (teste de independência: qui-quadrado = 2,056, p = 0,152), mas nos permite pensar que a qualidade dos laços dos informantes dos grupos mais jovens seja mais fraca do que as dos informantes dos grupos mais idosos, razão pela qual eles e os seus contatos estão mais sujeitos às mudanças, a palatalizar, no que tange ao presente estudo.

Os resultados da análise não confirmam nossa hipótese. Quanto à posição que ocupam na rede, os inovadores, informantes jovens de Antônio Prado, não são periféricos. Quanto à qualidade de seus laços, verifica-se o predomínio dos fracos, embora estatisticamente não significativo. Antes do que frustrar nossas expectativas, talvez a análise tenha nos revelado a real motivação para a estabilização da regra na comunidade. Não sendo periféricos, os inovadores têm maior potencial difusor da regra na rede; no entanto, a qualidade dos laços que os conectam aos demais membros é fraca, o que refreia a difusão, dando origem à moderada palatalização.

Pelo que se verificou então, a plexidade (qualidade ou conteúdo dos laços) tem impactos no progresso da regra. A plexidade relaciona-se às práticas sociais realizadas pelos membros. Que práticas sociais são essas?

3.3. RCI-RS e identidade: a palatalização entre o local e o global

A variante palatalizada é inovadora em comunidades da antiga região colonial italiana do Rio Grande do Sul, como Antônio Prado e Flores da Cunha. Na realização das entrevistas sociolinguísticas e em atividades de observação participante, não se percebeu qualquer associação da variante palatalizada com valores positivos como bonito, moderno, correto. Tampouco negativos. O que se percebeu foi uma identificação dessa variante com o "de fora", com o que não é local. Em nossa interação com os informantes, foi nítido seu esforço de, num primeiro momento, produzir algumas formas palatalizadas, muitas das quais abandonadas ao longo da conversa. Outros, principalmente os mais velhos, não palatalizavam. Mas sugeriam, ou até afirmavam claramente, que "aqui não se fala assim", mesmo que, inconscientemente, tivessem palatalizado vez ou outra. Isso mostra que essa variante, saliente na região, pode ser usada como um recurso identitário, para aproximar os falantes dos forasteiros ou, fora dos limites da comunidade, no âmbito aqui denominado global, para encobrir sua identidade.

Essa utilização das variantes como recursos linguísticos fora do local ocorrerá se houver mobilidade territorial. Os indivíduos fora da comunidade estão

sujeitos à exposição a outros padrões de fala, cujas características podem ser incorporadas à sua. No que se refere a Antônio Prado, Flores da Cunha e municípios vizinhos de mesma origem étnica, a italiana, o global e o local necessitam ser entendidos relativamente à antiga região colonial italiana do Rio Grande do Sul. As fronteiras de boa parte dos municípios dessa região são constantemente redesenhadas pelas relações (econômico-culturais) que estabelecem com municípios vizinhos maiores, destacando-se aí Caxias do Sul, e com outras localidades. Essas repercutem diretamente nas práticas sociais individuais. Por exemplo, para estudar, os habitantes mais jovens deslocam-se diariamente a outras comunidades. Por outro lado, mesmo com os setores do comércio e da indústria relativamente bem desenvolvidos, uma parcela da população dedica-se às práticas agropecuárias que mantêm os indivíduos na zona rural dos municípios. Esse conjunto de práticas socioeconômicas tem sido relacionado à vocação para o trabalho e ao empreendedorismo dos imigrantes italianos, traços celebrados em festas comunitárias como a Noite Italiana (Antônio Prado), a Festa da Vindima (Flores da Cunha) e a Festa da Uva (Caxias do Sul). Estruturas sociais tradicionais, como a familiar, ainda orientam as práticas individuais, o que se verifica na rede social. Embora os informantes estabeleçam relações supraterritoriais em algumas de suas práticas, convivem na comunidade conforme os velhos padrões da família patriarcal, o que denota, em termos de identidade, uma orientação para o local e acaba refreando a expansão maciça de elementos globais sobre os traços locais. Em termos linguísticos, isso corresponde a uma situação de aparente transitoriedade: há variação, mas moderada. Além desses aspectos, é preciso considerar que a antiga região colonial italiana do Rio Grande do Sul, situada na América Latina e no Brasil, já nasceu de processos históricos globalizadores (MENZ, 2009; KÜHN, 2007; GIRON, 1992), a que se devem as dificuldades de criação de um sentimento e de uma ideologia que pudessem ser rotulados nacionais (OLIVEN, 1992; SEYFERTH, 2000). Ideologicamente, quando a antiga região colonial italiana do Rio Grande do Sul começou a apresentar índices significativos de crescimento e desenvolvimento, já na segunda metade do século XX, a antiga tradição italiana foi reconstruída. Como consequência, a incorporação globalizadora não tem sido tão rápida. O local é relativamente desenvolvido, apegado ao passado e às tradições da colonização. Isso reforça valores ligados ao mundo do trabalho, da religião, da família e fornece recursos, entre eles os linguísticos, para a afirmação da uma identidade local. É o que faz emergir padrões moderados de mudança nos comportamentos sociais e nos usos linguísticos na antiga região, refreando a expansão da palatalização variável da capital a essa região do interior, o que afeta os jovens em suas práticas sociais diárias.

3.4. Os jovens e a palatalização variável na antiga região colonial italiana do Rio Grande do Sul

À primeira vista, os jovens da antiga região colonial italiana do Rio Grande do Sul não diferem de qualquer outro jovem brasileiro. Considerando o estado do Rio Grande do Sul em termos de vestimentas, por exemplo, assemelham-se à maioria dos jovens gaúchos. Mas a observação mais sistemática de suas práticas sociais e a consideração aos assuntos de que falam, às questões sobre que debatem, revelam um localismo peculiar. Esse localismo, em parte, explica as vinculações de suas práticas com as tradições da imigração italiana, noutra parte, a necessidade/desejo de inovar e rever essas vinculações.

Talvez reforçada pelas festas locais e pelo turismo, que celebra e comercializa as raízes italianas, há uma consciência étnica que serve de explicação, para o próprio jovem, da razão de realizar certas práticas, como se pode captar de uma afirmação assim:

> (...) a maioria (dos jovens) aqui da cidade tem descendência italiana. (Tu) sabe que descendência italiana sempre tem a matriarca e o patriarca. Eu acho que a maioria sai, assim, bastante "família" e sempre tem que jantar junto com o pai e com a mãe, salvo exceções. (...) Chega essa idade assim e "ah, tenho que casar, tenho que ter minha família". (FC; M. de O., 24 anos, masculino, zona urbana)

A vida em família não parece ser algo que o jovem gostaria de evitar, tampouco um real desejo seu, mas algo que acontece com ele. O jovem nem se rebela contra a família, nem adere incondicionalmente a ela, o que pode estar na base das graduais e pequenas mudanças diárias por ele promovidas, embora as tradições sejam seguidas, inclusive nos momentos de lazer:

> Aqui (se) segue...um pouco de tradição da cidade, que no caso é, no domingo, sair e dar voltas ao redor da praça. É uma coisa que a cidade pequena tem. Ficar parado olhado o movimento passar. (FC; M. de O., 24 anos, masculino, zona urbana)

Essa fala não revela propriamente uma opção do jovem, mas uma consequência das tradições e, mais importante, do fato de a cidade não oferecer opções de lazer como teatro, cinema, *shopping center*, comentário que soa como queixa na fala de muitos deles e que é razão para buscarem, em outras cidades, diferentes formas de distração:

> Eu gostaria de viver em Caxias... Acho que desde pequena... uma vez eu fui na Festa da Uva lá e nossa! ... sempre gostei de Caxias, tanto pequena quanto adolescente, sempre quis

> ir pra Caxias, mas minha mãe não deixava e, nossa... eu ia no cinema, ficava feliz, sempre gostei da cidade (C. S., 18 anos, feminino, zona urbana)

Deslocar-se a outras localidades, dessa forma, parece fazer parte de ser jovem na antiga região colonial italiana do Rio Grande do Sul, inclusive para cursar faculdade, mas em um roteiro de ida e volta: o jovem realiza práticas fora da comunidade, mas retorna a ela. Em termos linguísticos, tem contato com outros padrões de fala e pode, nessa circunstância, sentir-se pressionado a evitar formas reveladoras de sua identidade local, como não palatalizar. É possível então que, buscando sintonia com a fala do outro, abra mão, momentaneamente, de marcas locais. Depois, esse mesmo jovem retorna à comunidade e lá, volta a orientar-se pela identidade e valores locais, embora não realize práticas sociais exatamente da mesma forma que os indivíduos mais velhos. Sobre sair da comunidade e retornar a ela e também sobre mudanças geracionais, observe-se a afirmação de outra jovem sobre trabalho:

> A gurizada daqui não quer mais trabalhar na colônia. (...) Vão pra Caxias estudar. Todos fazem faculdade, a grande maioria. Aí, então, ninguém mais quer trabalhar na colônia como acontecia antigamente, as famílias eram numerosas, as pessoas ficavam na colônia e não estudavam, né. (...) Eu acho que mais pessoas continuam morando aqui e vão e voltam, do que se mudam. São poucos que se mudam. (FC; C. M., 23 anos, zona rural)

A jovem começa afirmando que a gurizada não quer mais trabalhar na terra, nas propriedades rurais, junto à família. A suposição imediata seria a de que os jovens da zona rural desejariam migrar para a cidade, mas não é o que vai na última afirmação da jovem, repetida a seguir: "Eu acho que mais pessoas continuam morando aqui e vão e voltam, do que se mudam. São poucos que se mudam". É o tipo de trabalho que jovens de zona rural criticam: muito duro, segundo alguns. Mas não criticam o lugar onde moram, que não abandonam, se possível. É o que observamos em campo. Nas propriedades rurais onde se cultiva uva, por exemplo, setor bastante lucrativo para os pequenos proprietários da região – jovens que, na adolescência, planejaram abandonar a zona rural afirmam ter se dado conta, alguns anos depois, de que a viticultura lhes daria um retorno financeiro interessante. E de fato trabalham na terra, acompanhando seus pais.

Sobre práticas linguísticas, em específico, são raras as afirmações a respeito nas entrevistas sociolinguísticas, como é de se esperar. Mas veja-se uma interessante, de uma jovem que falava sobre padrões familiares:

> Assim: começa sempre pelos avós, então, a criação dos meus tios, né, era uma e veio vindo ... e a gente sempre pegava (...) o próprio falar: tu fala "erado", e quem convive contigo que fala "erado" vai falar sempre "erado". (FC; C. P., 18 anos, feminino, zona urbana)

A jovem tem ciência de que, ao falarmos, tendemos a reproduzir práticas linguísticas, algumas delas desprestigiadas fora da comunidade. Essa afirmação revela, em nosso entender, uma das motivações para os mais jovens, gradualmente, palatalizarem, embora a não-palatalização emerja com bastante naturalidade na fala de alguns.

Na antiga região, as práticas religiosas, quase exclusivamente católicas, têm relevo. Embora nas entrevistas sociolinguísticas esse tópico não tenha rendido uma boa conversa, temos observado o envolvimento de jovens em práticas religiosas, principalmente as festivas. Na celebração de *Corpus Christi* em Flores da Cunha, por exemplo, que envolve uma missa na igreja matriz e, em seguida, a procissão do Senhor Morto sobre tapetes de serragem, constatamos a presença de número significativo de jovens. Não pareciam estar lá forçados, pelo contrário, sua participação demonstrou-se fervorosa. Constatamos também que muitos tapetes haviam sido confeccionados por grupos de jovens – juventude católica, escoteiros, organizações antidrogas, grêmios desportivos.

Percebemos nos jovens os *jeans*, tênis, jaquetas, bonés e celulares que os acompanhariam a qualquer outro lugar do Rio Grande do Sul. Um pouco mais cuidados e "arrumadinhos", é verdade, mas todos esses itens estavam lá no visual, numa demonstração de que a prática é local, mas os artefatos são supralocais.

Em termos de práticas social e cultural local, vemos no estudo etnográfico indícios de um hibridismo, em que aspectos tradicionais e inovadores convivem, em que mudanças linguísticas (entre outras) ocorrem, mesmo que lentamente, acompanhando mudanças sociais.

CONSIDERAÇÕES FINAIS

Na pesquisa sociolinguística, a análise de redes e práticas sociais pode esclarecer o papel das ligações entre as pessoas, da maior ou menor coesão dos grupos, da pressão dos pares e das identidades locais na variação e mudança linguística.

A abordagem sobre as noções teóricas e a retomada de trabalhos sobre a palatalização variável de /t/ e /d/ no português do sul do Brasil, possibilitaram discutir, propor e ilustrar procedimentos metodológicos passíveis de emprego na análise da variação linguística e práticas sociais. Mostraram que o tratamento quantitativo fornece um importante diagnóstico das possíveis motivações para os processos variáveis. Mas é apenas dando um passo além do cálculo estatístico, ou seja, buscando interpretar os resultados através de microanálise, pela investigação de práticas diárias em comunidade e de sua coesão (em rede), que se pode saber o que está por trás desses números, em especial, sobre os sujeitos que falam.

REFERÊNCIAS

BATTISTI, E. Variação, mudança fônica e identidade: A implementação da palatalização de /t/ e /d/ no português falado na antiga região colonial italiana do Rio Grande do Sul. *Diadorim*. Rio de Janeiro: v. 8, p. 103-123, 2011.

BATTISTI, E.; DORNELLES FILHO, A.A.; LUCAS, J.I.P.; BOVO, N.M.P. Palatalização das oclusivas alveolares e a rede social dos informantes. *Revista virtual de estudos da linguagem – Re*VEL, v.5, n.9, agosto 2007. Disponível em: <www.revel.inf.br>. Acesso em: 27 fev.2008.

_____. Palatalização das oclusivas alveolares e a dimensão subjetiva da variação. *Caderno de Pesquisas em Linguística* – Variação no Português Brasileiro. Porto Alegre, v.3, n.1, 2007b.

_____. *Dental stops palatalization a social practice*. Talk presented at SS17, Free University of Amsterdam, Amsterdam, 5 April 2008.

BATTISTI, E.; DORNELLES FILHO, A.A.; LUCAS, J.I.P. A implementação da palatalização das oclusivas alveolares no português brasileiro: rede social e ideologia. In: VI *Congresso Internacional da* ABRALIN, *2009*, *João Pessoa*. ABRALIN - VI *Congresso Internacional:* ANAIS. João Pessoa: Ideia, 2009. p. 1249-1258.

BATTISTI, E.; GUZZO, N.B. Palatalização das oclusivas alveolares: O caso de Chapecó. In: BISOL, L.; COLLISCHONN, G. (Orgs.) *Português no sul do Brasil:* Variação fonológica. Porto Alegre: EDIPUCRS, 2009. p.114-140.

BATTISTI, E.; LUCAS, J.I.P. Língua, redes e práticas sociais. In: BATTISTI, E.; CHAVES, F.G.L. (Orgs.) *Cultura Regional:* Língua história, literatura 2. Caxias do Sul: EDUCS, 2006. p.113-131.

BATTISTI, E. ; MARTINS, L.B. A realização variável de vibrante simples em lugar de múltipla no português falado em Flores da Cunha (RS): Mudanças sociais e linguusticas. *Cadernos do* IL, Rio Grande do Sul, v. 42, p. 146-158, 2011.

BLAKE, R.; JOSEY, M. The /ay/ diphthong in Martha's Vineyard community: what can we say 40 years after Labov? *Language in Society*, Cambridge, v.4, n. 32, p.451-485, 2003.

BOAVENTURA NETTO, P. O. *Grafos*: teoria, modelos, algoritmos. São Paulo: Ed. Edgard Blücher, 1996.

BONNEWITZ, P. *Primeiras lições sobre a sociologia de P. Bourdieu*. Petrópolis: Vozes, 2003.

BOURDIEU, P. *Outline of a theory of practice*. Cambridge: Cambridge University Press, 1977.

CASTELLS, M. *A sociedade em rede*. Rio de Janeiro: Paz e Terra, 1999.

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

EVANS, B. The role of social network in the acquisition of local dialect norms by Appalachian migrants in Ypsilanti, Michigan. *Language Variation and Change*, Cambridge, 16 v, p.153-167, julho 2004.

FRUCHTERMAN, T. M. J.; REINGOLD, E. Graph drawing by force-directed placement. *Software-Practice and Experience*, 21 v, n.11, p. 1129-1263, novembro 1991.

GERHARDT, G. J. L.; CORSO, G.; LEMKE, N. *Network clustering coefficient approach for* DNA *sequences (pre-print)*, 2005.

GIRON, L. S. A imigração italiana no Rio Grande do Sul: fatores determinantes. In: DACANAL, J. H. (Org.). Rio Grande do Sul: imigração & colonização. 3.ed. Porto Alegre: Mercado Aberto, 1996.

KÜHN, F. Uma fronteira do Império: O sul da América portuguesa na primeira metade do século XVIII. *Anais de História de Além-Mar*, v.VIII, 2007. p.103-121.

LABOV, W. *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press, 1972.

_____. *Principles of linguistic change*: *internal factors*. Malden/Oxford: Blackwell, 1994.

_____. *Principles of linguistic change: cognitive and cultural factors*. Malden/Oxford/Sussex: Wiley-Blackwell, 2010.

LI WEI. Network analysis. In: GOEBL, H.; NELDE, P.; ZDENEK, S.; WOELCK, W. (eds.). *Contact linguistics: a handbook of contemporary research*. Berlin: de Gruyter, 1996.

MATHEWS, J. H. *Numerical methods for mathematics*, *science and engineering*. Prentice Hall, Englewood Cliffs, 1992.

MENZ, M.M. *Entre impérios*: Formação do Rio Grande na crise do sistema colonial português. São Paulo: Alameda, 2009.

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

_____. Social networks. In: CHAMBERS, J.K.; TRUDGILL, P.; SCHILLING-ESTES, N. (eds.) *The Handbook of Language Variation and Change*. Malden/Oxford: Blackwell, p.549-572, 2002.

MILROY, L.; MILROY, J. Linguistic change, social network and speaker innovation. In: *Journal of Linguistics*, Cambridge: Cambridge University Press, v. 21, p.339-384, 1985.

_____. Social networks and social class: Toward an integrated sociolinguistic model. *Language in Society*, Cambridge: Cambridge University Press, v. 21, p.1-26, 1992.

NOLL, V. *O português brasileiro:* formação e contrastes. São Paulo: Globo, 2008.

OLIVEN, R.G. *A parte e o todo*: A diversidade cultural no Brasil-Nação. Petrópolis: Vozes, 1992.

SEYFERTH, G. Identidade nacional, diferenças regionais, integração étnica e a questão imigratória no Brasil. In: ZARUR, G. de C.L. *Região e nação na América Latina*. Brasília: Editora da UnB, p.81-100, 2000.

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

CAPÍTULO 8

REDES SOCIAIS, VARIAÇÃO LINGUÍSTICA E POLIDEZ: PROCEDIMENTOS DE COLETA DE DADOS

Andréia Silva Araujo
Kelly Carine dos Santos
Raquel Meister Ko. Freitag

INTRODUÇÃO

Como Elisa Battisti mostrou no capítulo anterior, as redes sociais vêm sendo utilizadas nos estudos variacionistas a fim de contribuir para a análise dos processos de variação e mudança linguísticas, uma vez que, com essa metodologia, torna-se possível realizar uma análise com base na frequência e qualidade da interação dos membros constituintes das redes, abrindo espaço para um estudo voltado ao campo da pragmática. Entrevistas sociolinguísticas têm sido fonte produtiva para a realização de descrição linguística no português; no entanto, não possibilitam realizar uma análise voltada para os papéis sociopessoais dos interlocutores envolvidos (entrevistador-entrevistado) (FREITAG, 2010; 2012), o que é essencial para captar os efeitos dos valores de polidez nos usos linguísticos.

A polidez é uma estratégia linguística utilizada com o objetivo de evitar conflitos na interação verbal. Segundo Brown e Levinson (2011), trata-se de uma estratégia para preservarmos a nossa face e a do outro com o intuito de estabelecer uma comunicação econômica e eficaz, sem atritos. O valor de polidez emerge em contextos específicos, com fatores fortemente correlacionados: do ponto de

vista pragmático, a distância social, as relações de poder/poder relativo e o custo da imposição, são fatores fortemente envolvidos na avaliação de quais estratégias linguísticas são mais ou menos polidas (BROWN; LEVINSON, 2011); e do ponto de vista sociolinguístico, a relação entre sexo/gênero dos interlocutores mostra-se significativa. Para constituir uma amostra de fala que capte os efeitos de polidez envolvidos no processo interacional, faz-se necessário desenvolver uma metodologia que permita apreender esses fatores.

1. INTERAÇÕES CONDUZIDAS: COMUNIDADES DE PRÁTICAS E REDES SOCIAIS

Para controlar a correlação entre os graus de proximidade, relações de poder, custo da imposição, aspectos pragmáticos e o sexo/gênero – variável sociolinguística clássica – é preciso delinear uma estratégia de coleta que considere uma unidade de análise – comunidade de prática – e uma proposta de hierarquização – redes sociais.

1.1. A comunidade de prática em foco

Uma comunidade de prática é um agrupamento de pessoas que se engajam em um empreendimento comum e é durante esta atividade conjunta que as práticas emergem – "o modo de fazer as coisas, modos de falar, crenças, valores, relações de poder" – (ECKERT; MCCONNEL-GINET, 2010, p. 102). As autoras explicitam que uma comunidade de prática pode ser representada "por pessoas trabalhando juntas em uma fábrica, *habitués* de um bar, companheiros de brincadeira em uma vizinhança, a família nuclear, parceiros policiais e seu etnógrafo, a Suprema Corte etc.". E ressaltam que:

> Comunidades de prática podem ser grandes ou pequenas, intensas ou difusas; elas nascem e morrem, podem sobreviver a muitas mudanças de membros e podem estar intimamente articuladas a outras comunidades. As pessoas participam de múltiplas comunidades de prática, e a identidade individual é baseada nesta participação. Em lugar de conceber o indivíduo como uma entidade à parte, pairando sobre o espaço social, ou como um ponto em uma rede, ou como membro de um conjunto específico ou de um conjunto de grupos, ou como um amontoado de características sociais, precisamos enfocar as comunidades de prática. Tal foco possibilita-nos ver o indivíduo como agente articulador de uma variedade de formas de participação em múltiplas comunidades de prática. (ECKERT; MCCONNEL-GINET, 2010, p. 102-103)

Essa perspectiva de comunidade tem sido tomada por sociolinguistas, que estudam a variação em uma dimensão estilística, por objetivarem captar com mais detalhes a dinâmica do valor social das variáveis (FREITAG; et al., 2012) e, assim, observarem como ocorre a construção da identidade do indivíduo e a construção do significado social. Estudos nessa perspectiva têm sido chamados de terceira onda da sociolinguística. A terceira onda incorpora postulados dos estudos da primeira e a segunda, mas com um diferencial: o foco passa da comunidade de fala para a comunidade de prática (FREITAG; et al., 2012).

Para constituirmos a amostra "Rede social de informantes universitários", escolhemos uma comunidade de prática pertencente à cidade de Itabaiana/SE. Esta cidade localiza-se no agreste central do Estado de Sergipe, a 58 km da capital Aracaju. O município possui uma área de 336.693 km² e tem uma população estimada em 91.873 habitantes. A Figura 1 a seguir destaca a localização geográfica da cidade de Itabaiana no mapa de Sergipe.

Figura 1 – Localização de Itabaiana/SE no mapa de sergipe. Fonte: Wikipédia

Em decorrência do programa do Governo Federal de expansão e interiorização da educação superior no Brasil, a cidade de Itabaiana recebeu um *campus* universitário – Universidade Federal de Sergipe, *campus* Prof. Alberto Carvalho. Suas atividades foram iniciadas no dia 14 de agosto de 2006 e há, atualmente, dez cursos em funcionamento, entre os quais sete são de licenciatura. O *campus* recebe cerca de 2.500 estudantes diariamente provenientes da cidade de Itabaiana e das circunvizinhas.

A instalação do *campus* foi muito importante para os habitantes dessa região, por aumentar a possibilidade de estes terem acesso ao nível superior. Segundo Freitag (2012, p. 932), "ser universitário é uma conquista familiar da maioria: pesa a responsabilidade de ser o primeiro universitário em uma família de pais

que não tiveram a oportunidade de ter acesso à escolarização". Os estudantes passam pelo menos quatro horas diárias no ambiente universitário, desenvolvendo atividades, compartilhando valores e conhecimentos. Há estudantes que, seja por morarem distante, seja por morarem em outra cidade, vão para a universidade em ônibus escolares ou particulares e durante o trajeto, estabelecem contato uns com os outros. Esse engajamento social que há entre os universitários do *campus* de Itabaiana nos permite defini-lo como uma comunidade de prática, nos termos que propõem Eckert e McConnell-Ginet (2010), porque há um conjunto de pessoas agregadas para aprender, construir e fazer a gestão do conhecimento.

1.2. Redes sociais

Para a construção de um modelo metodológico de constituição de banco de dados de fala que capte nuanças de polidez, partimos da hipótese de que não podemos ter a figura do entrevistador para conduzir o tópico como ocorre nas entrevistas sociolinguísticas (nos moldes canônicos), pois o entrevistador pode influenciar o uso linguístico do entrevistado, ocasionando o que tem sido denominado de efeito Rutledge[1]. Além disso, nas entrevistas sociolinguísticas não é possível realizar uma "análise mais acurada dos papéis sociopessoais do entrevistador e sua relação com o entrevistado" (FREITAG, 2012, p. 295), o que é essencial para captar o valor de polidez. Para tanto, precisamos que os próprios informantes conduzam o tópico na interação para que assim possamos controlar o sexo, a distância social/grau de proximidade, o custo da imposição e as relações de poder estabelecidas por estes. Dessa forma, o pesquisador não participa da interação, para minimizar a influência nos dados coletados. Chamaremos essa situação de fala, em que os próprios informantes conduzem o tópico, de interações conduzidas, um procedimento metodológico de coleta de dados que se assemelha ao grupo focal.

Entre os fatores considerados, está a distância social entre os informantes, de ordem pragmática. Os usos linguísticos de um indivíduo estão fortemente correlacionados a essa distância. Isso significa dizer que, se um indivíduo tem um grau de proximidade forte com um interlocutor e fraco com outro, seu comportamento linguístico na interação com cada um deles é, provavelmente, diferente em decorrência do tipo de relacionamento. Portanto, o controle dessa variável

1 O efeito Rutledge é um conceito decorrente da reanálise do estudo de Montgomery (1998) em que se constatou que os resultados obtidos quanto à distribuição de *might could* em função do sexo/gênero foram influenciados por uma entrevistadora, Barbara Rutledge, que sugeria a resposta com a forma *might could*. (cf. FREITAG, 2012).

nos permite verificar se de fato os diferentes usos linguísticos são decorrentes do grau de proximidade existente entre os informantes. Mas como constituir um banco de dados controlando essa variável?

Buscamos respaldo nos modelos que consideram as redes sociais para observar/controlar as relações existentes entre seus membros. Tal teoria foi desenvolvida por antropólogos nas décadas de 1960 e 1970 e introduzida na Sociolinguística por Lesley Milroy, a partir da década de 1980. Os sociolinguistas utilizam essa teoria em suas análises "para verificar o papel do falante na inovação linguística (ou o bloqueio dela)" (BATTISTI, 2008, p.2).

Entende-se por rede social o conjunto de atores/pessoas que têm relações entre si, sejam elas por laços fortes (grau de proximidade alta) ou fracos (grau de proximidade baixa). Para controlar o grau de proximidade, é necessário focar em uma comunidade menor, para observar as redes sociais pessoais constituídas. No entanto, para a construção de um modelo metodológico que capte nuanças de polidez, não é necessário identificar as diversas redes sociais das quais participam o informante e todos os tipos de intensidades de grau de proximidade, já que para controlar os fatores dimensão social, relações de poder, custo da imposição e sexo, cada informante terá que despender uma grande quantidade de tempo, cerca de 5 horas, para a realização das gravações das interações. Pode-se focalizar apenas em uma comunidade de prática e escolher um tipo de rede de relacionamento pessoal existentes. Por exemplo, dentro da comunidade de prática escolar, existem várias redes de relacionamentos, tais como: aluno-aluno, aluno-professor, aluno-funcionário, professor-funcionário, professor-professor, funcionário-funcionário. Para a formação da rede social, o primeiro passo para realizar a coleta é delimitar a comunidade de prática e escolher o tipo de rede a ser analisada.

O pesquisador pode identificar uma rede observando quem interage com quem, em uma dada comunidade e como ou por que eles estão interagindo, ou ainda perguntando quem são os melhores amigos, os com quem eles conversaram ontem, para que assim as pessoas definam suas próprias redes.

Essa relação de contato com os outros ainda pode ser vista como uma teia infinita de laços, que se estendem através de toda a sociedade interligando os seus membros (MILROY, 2002) com laços de primeira ordem, formado por pessoas que diariamente estão interagindo, e laços de segunda ordem, a partir dos quais as pessoas se interligam indiretamente.

Outro aspecto observável nas redes diz respeito à sua densidade e multiplexidade, como mostrou Battisti no capítulo anterior. Segundo Meyerhoff (2006), redes de baixa densidade tornam os indivíduos mais abertos à mudança, pois os laços que eles terão com outras redes irão contribuir para que utilizem inovações que adquiriram. Da mesma forma, Milroy (2002) acredita que as redes

constituídas de laços fortes (densa e multiplexa) contribuem para que a variedade linguística da comunidade resista a mudanças linguísticas. Os indivíduos podem participar de grupos diferentes e os laços fortes e fracos tendem a conectá-los, ligados em graus diferentes, já que os membros de uma rede podem se conhecer e se relacionar a partir de diversos graus de intimidade.

2. DELINEAMENTO DA AMOSTRA

O primeiro passo para a representação da rede social é a escolha da comunidade de prática e o tipo de relacionamento pessoal que se quer analisar, conforme ressaltamos. Feito isso, devem ser identificados os informantes que fazem parte da rede social escolhida. Não é necessário para a constituição do banco de dados que todos que fazem parte sejam considerados. Conseguimos contar com oito informantes, quatro homens e quatro mulheres, para a constituição da amostra. O Quadro 1 traz a distribuição e algumas informações sociais dos informantes que participaram da amostra.

Sexo/gênero	Informante	Idade	Curso/período	Cidade
Feminino	D. C.	28	Geografia/8º	Itabaiana
	A. G.	25	Geografia/8º	Campo do Brito
	J. S.	19	Pedagogia/2º	Frei Paulo
	L. R.	21	Pedagogia/8º	Frei Paulo
Masculino	D. S.	21	Geografia/8º	Itabaiana
	D. M.	24	Geografia/8º	Itabaiana
	W. S.	19	Ciências Cont./2º	Frei Paulo
	C. A.	30	Administração/9º	Frei Paulo

Quadro 1 – Distribuição e dados sociais dos informantes em rede social pessoal.

Os oito informantes formam dois grupos – cada um com duas mulheres e dois homens – em que aqueles que pertencem a um grupo têm relações de proximidade entre si, mas não com os informantes pertencentes ao outro. (Figura 2)

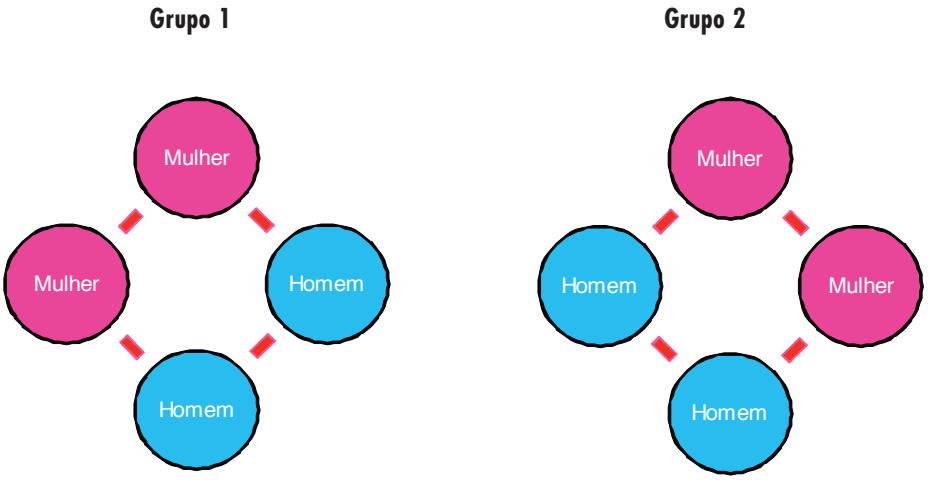


Figura 2 – Arranjo dos informantes

A distância social/grau de proximidade entre os informantes é controlada por meio da frequência com que interagem. Para mensurar a distância social/grau de proximidade entre os informantes, nos baseamos na proposta de controle de Blake e Josey (2003) e de Oushiro (2011). O controle deste fator foi estipulado a partir de uma escala de 1-5, que vai do grau máximo (grau 1) ao grau mínimo de proximidade (grau 5) entre os informantes (Quadro 2).

Grau 1 – Bastante próximo. Os informantes possuem laços fortes (amizade, parentesco, colega de trabalho ou escola etc.) e interagem diariamente;
Grau 2 – Próximo. Os informantes interagem frequentemente, mas não possuem laços fortes;
Grau 3 – Próximo. Os informantes não interagem frequentemente e não possuem laços fortes;
Grau 4 – Neutro. Os informantes 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.

Quadro 2 – Escala de gradação para o controle da distância social entre os informantes da rede social.

Partimos da premissa de que os usos linguísticos dos informantes variem de acordo com o grau de proximidade existente entre eles: i) quanto mais forte for o grau de proximidade entre os informantes, menor será o número de ocorrências de estratégias de polidez (menos polido); e ii) quanto mais fraco for o grau

proximidade entre os informantes, maior será o número de ocorrências de estratégias de polidez (mais polido).

Quanto mais variáveis controlamos, maior é o número de informante e o tempo que cada um precisa dispor para a gravação das interações; por isso, focamos nos extremos: grau 1 e grau 5. O controle dos extremos da escala de gradação é suficiente para se verificar os efeitos de polidez decorrentes do grau de proximidade entre os interlocutores. A interação dos informantes dentro e entre os grupos deve ocorrer da seguinte forma: cada informante deve interagir com um homem e uma mulher com os quais tenha grau 1 de proximidade e com um homem e uma mulher com os quais tenha grau 5. Dessa forma, ao controlar o grau de proximidade entre informantes, controlamos também a variável sexo/gênero. O controle dessa variável desdobra-se, portanto, em quatro fatores:

i) feminino → masculino;
ii) feminino → feminino;
iii) masculino → feminino;
iv) masculino → masculino.

Para possibilitar o controle da variável pragmática relações de poder, cada informante interagiu duas vezes com o mesmo interlocutor. Em uma das interações, um dos informantes conduziu o tópico e, na outra, trocaram de papéis. Por exemplo: em uma interação o falante 1 conduz o tópico com o falante 2; e na outra inverte-se a situação, o falante 2 conduz o tópico na interação com o falante 1. Tal troca de papéis sociopessoais nos permite, portanto, controlar as relações de poder envolvidas a partir de quem está com o domínio do tópico na interação. O custo da imposição é controlado por meio do tipo de assunto que é introduzido, situações que vão da aparente neutralidade às que envolvem a preservação das faces negativa e positiva. A classificação do tipo de assunto introduzido na interação pode ser vista dentro de um *continuum* que vai do [- impositivo] ao [+ impositivo]. (Figura 3)

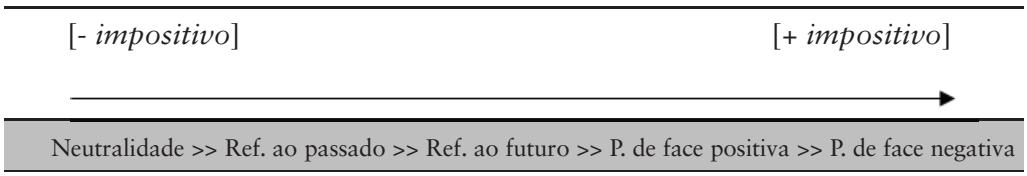


Figura 3 – *Continuum* do tipo de assunto quanto ao custo da imposição.

As categorias controladas possibilitam o estabelecimento das seguintes relações/conexões na rede social de informantes:

- cada informante mulher deve conduzir o tópico da interação com uma informante de grau 1 de proximidade, que também deve conduzi-lo em uma segunda interação;
- cada informante mulher deve conduzir o tópico da interação com um informante de grau 1 de proximidade, que também deve conduzi-lo em uma segunda interação;
- cada informante homem deve conduzir o tópico da interação com uma informante de grau 1 de proximidade, que também deve conduzi-lo em uma segunda interação;
- cada informante homem deve conduzir o tópico da interação com um informante de grau 1 de proximidade, que também deve conduzi-lo em uma segunda interação;
- cada informante mulher deve conduzir o tópico da interação com uma informante de grau 5 de proximidade, que também deve conduzi-lo em uma segunda interação;
- cada informante mulher deve conduzir o tópico da interação com um informante de grau 5 de proximidade, que também deve conduzi-lo em uma segunda interação;
- cada informante homem deve conduzir o tópico da interação com uma informante de grau 5 de proximidade, que também deve conduzi-lo em uma segunda interação;
- cada informante homem deve conduzir o tópico da interação com um informante de grau 5 de proximidade, que também deve conduzi-lo em uma segunda interação.

Cada informante interagiu com 4 pessoas diferentes (um homem e uma mulher, próximos dele; um homem e uma mulher, distantes dele) duas vezes, totalizando 32 interações conduzidas. Na Figura 4 delineamos a rede social formada pelas conexões estabelecidas entre os informantes na amostra.

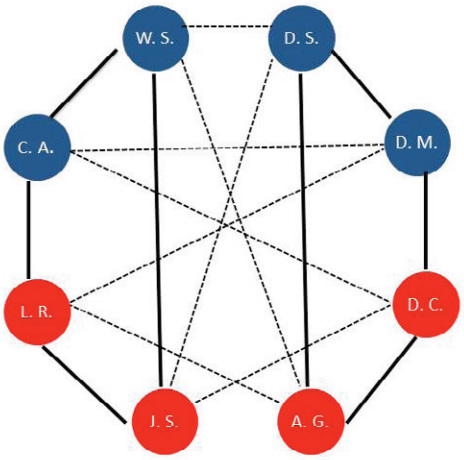


Figura 4 – Representação da rede social dos oito informantes universitários.

Os informantes do primeiro grupo (D.S., D. M., D. C. e A. G.) estão conectados por um laço forte, representado pela linha contínua, já que estes mantêm contanto diário, enquanto que todos eles mantêm laços fracos, representados pelas linhas tracejadas, com os membros do segundo grupo (W. S., C. A., L. R. e J. S.). Da mesma forma, os membros do segundo grupo mantêm laços fortes entre si, e laços fracos com os componentes do segundo grupo.

Na Figura 4, os representantes do sexo/gênero masculinos estão representados pela cor azul, e os representantes do sexo/gênero feminino pela cor vermelha. A cada conexão de dois informantes, tivemos duas interações (em um primeiro momento um informante conduziu o tópico na interação, e em um segundo momento quem conduziu o tópico foi o que interagiu), totalizando 32 interações conduzidas.

3. GRAVAÇÃO DAS INTERAÇÕES

A fim de minimizar os efeitos do paradoxo do observador nas coletas, não houve entrevistador e nem roteiro de perguntas, foram os próprios informantes que selecionaram o tópico da interação a partir de situações descritas em cartões. No momento da gravação, disponibilizamos 50 cartões, com temas diversos sobre situações que recobrem:

- **neutralidade:**
  "O aumento do uso de redes sociais online é exorbitante. Pessoas de todas as idades têm aderido a esse sistema de comunicação.";

- **referência ao passado:**
  "As intrigas sempre estão presentes entre os irmãos. Sempre há aquele momento em que um olha para o outro e diz: "nunca mais fale comigo". Mas poucas horas depois eles já estão juntos novamente.";
- **referência ao futuro**
  "Pedro está terminando a graduação e está muito preocupado com a sua vida após a faculdade. Muitos são os planos.";
- **preservação de face positiva**
  "Diego e Bárbara foram aprovados no concurso de medicina da UFS através do sistema de cotas para alunos da rede pública. Ultimamente alguns alunos não cotistas se negam a desenvolver trabalhos acadêmicos com eles.";
- **preservação de face negativa**
  "A disfunção erétil, ou seja, a incapacidade de conseguir ereção satisfatória para o ato sexual, que pode ser ocasionada pela falta de desejo, pela ejaculação precoce ou retardada etc., traz insatisfação tanto para o homem quanto para a mulher."

Para a elaboração das situações descritas nos cartões, realizamos grupos focais com homens e mulheres universitários, da mesma comunidade, em que se solicitou que listassem: i) cinco temas ou mais que você conversaria com alguém na sala de espera de um consultório; ii) cinco temas ou mais relacionados à coisas positivas de sua infância; iii) cinco temas ou mais relevantes para um universitário se posicionar; iv) cinco temas ou mais que um universitário não deveria falar sobre, por ser universitário; v) cinco temas ou mais sobre os quais você conversaria com um homem desconhecido/uma mulher desconhecida; vi) cinco temas ou mais sobre os quais você não conversaria com um homem desconhecido/uma mulher desconhecida; e vii) cinco temas ou mais sobre os quais você conversaria com uma amiga íntima/amigo íntimo. A partir da recorrência das respostas dadas, selecionamos temas e elaboramos situações que vão da aparente neutralidade a situações que envolvem a preservação das faces positiva e negativa.

Para a coleta da interação, cada informante escolheu aleatoriamente 10 cartões (dois de cada tipo). A partir da situação descrita no cartão, o informante deveria identificar o tema abordado e conduzir a conversa com o seu interlocutor sobre este. Por exemplo, na situação descrita referente à preservação de face negativa, um dos temas abordado é a disfunção erétil. Trata-se de um tema bastante delicado para conversar com alguém, até mesmo entre pessoas que tenham um grau de proximidade alto. Ao abordar essa temática, o informante coloca a sua face positiva em risco e em evidência a negativa. Ao tentar preservar-se, o

informante pode recorrer às estratégias de polidez e, dessa forma, minimizar os custos da imposição. Então poderia, por exemplo, abordar o assunto de forma tangencial, sem precisar perguntar diretamente se o interlocutor já "brochou" alguma vez, e perguntar: o que o informante faria se soubesse que seu companheiro andou expondo a vida íntima do casal para os colegas, o que o informante acha da atitude das pessoas que saem expondo sua intimidade ou que saem falando que o seu companheiro "brochou", etc.

A fim de garantir uma condição de interação que se aproximasse ao máximo possível de uma situação real e espontânea de interação, foi adotado o seguinte protocolo para a realização da coleta:

- identificar o informante com perfil compatível para participar das interações;
- esclarecer ao informante sobre a finalidade da coleta;
- obter do informante a concordância em participar da interação, resguardando seu anonimato;
- planejar com o informante os encontros para a coleta de cada uma das oito interações de no mínimo 40 minutos;
- escolher um lugar bem silencioso para a gravação da interação;
- orientar ao informante sobre como proceder na interação com o seu interlocutor:
  - há cinquenta cartões coloridos, o informante deve escolher dois cartões de cada cor;
  - a partir da situação descrita no cartão, o informante deve identificar o tema abordado e conduzir a conversa com o seu interlocutor, até esgotar o assunto e passar ao tema do cartão seguinte.
- deixar o gravador ligado desde o início da interação;
- após a gravação da interação, solicitar que o informante preencha a ficha social, bem como o Termo de Consentimento Livre Esclarecido;
- após cada interação, preencher o campo sobre o grau de relação entre os informantes participantes.

Finalizada a coleta das interações conduzidas entre os informantes, procedeu-se ao processo de transcrição das interações. Adotamos as normas utilizadas pelo Grupo de Estudos em Linguagem Interação e Sociedade (GELINS). Organizamos o *corpus* e criamos um código de identificação para cada um dos informantes, resguardando assim, o anonimato. Seguidos estes procedimentos, constituímos a "Rede Social de Informantes Universitários" (amostra de interações conduzidas), que faz parte do Banco de Dados "Falares Sergipanos" (FREITAG, 2013), que segue duas linhas de coleta – a de comunidades de fala (estratificação homogeneizada) e a de comunidades de práticas (relações sociopessoais). Atendendo às

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

4. EXCERTOS DA AMOSTRA

A título de ilustração, vejamos um excerto de cada tipo de situação – neutralidade, referência ao passado, referência ao futuro, estratégia de polidez positiva, estratégia de polidez negativa – da amostra constituída. Dispomos primeiramente o comando da situação e em seguida apresentamos o dado produzido.

1. Situação de neutralidade
Comando:

Atualmente, o aumento da procura de remédios genéricos vem crescendo, tal comportamento é decorrente do seu valor, mas mesmo assim, muitas pessoas não optam por esse tipo de medicamento.

Excerto 1:

F1: (...) *sobre a questão dos remédios genéricos assim o que é que você acha você acha que ele tem o mesmo... aquela questão de valor né por ser mais barato do que o... o medicamento comum você acha que ele tem o mesmo... assim faz o mesmo efeito ou você prefere na hora de comprar um remédio ir mais pelo original não pelo genérico?*

F2: olha assim... eu mesmo desconheço... esse remédio genérico porque... eu não me lembro o dia que eu tomei um genérico... na minha vida... eu ouço muito falar gente que prefere mesmo tanto por causa do valor mas eu creio assim... que ele não vai ter a mesma... o mesmo poder de finalidade do que o normal... não tem... porque se tivesse eles não queriam fazer mais barato né? você vê porque se se eles realmente ti- é tivessem... fosse a mesma coisa em relação a... o seu poder né de cura o ( )... você vê que os outros iam deixar de existir mas não... existe tanto um quanto o outro né? (A.G.$_{cdt}$ W.S.$_{sdt}$ D$_{FM}$07)

2. Situação com referência ao passado
Comando:

Ana era uma criança muito traquina. Todas as vezes em que saía para algum lugar, aprontava alguma.

Excerto 2:

F1: *David quando você era assim mais novo criança dava muito trabalho a sua mãe seus pais?*
F2: nunca dei trabalho a meu pai... sempre fui um menino comportado... minha cara de no... de inocente...
F1: *quando sai pra algum lugar assim com eles pra passear pra ir pro cinema um teatro pra algum lugar assim você você aprontava muito ou sempre se comportava direitinho?*
F2: eu sempre me comportava... (A.G.$_{cdt}$ D.S.$_{sdt}$ P F$_{M}$ 05)

3. Situação com referência ao futuro
Comando:

Pedro está terminando a graduação e está muito preocupado com a sua vida após a faculdade. Muitos são os planos.

Excerto 3:

F1: *é... vamo lá... quais são as perspectivas que você vê ao terminar a universidade?... é ferro né... ((RISOS))*
F2: rapaz... ( ) sair daqui... tá meio difícil porque a gente já não ve concurso né na área da gente... já não tem quase... eu fico pensando mais em sair daqui pegar o... o diploma de de ensino superior e... tentar aí esses concursos de nível superior porque se for pra esperar pra concurso na área da gente velho... tá osso viu... a gente vê aí dez anos cincos anos doze anos pra passar um concurso que vem cinco seis vagas na área da gente... aí arriscando... porque tá osso pra... emprego... meio complicado velho... pra trabalhar no comércio aqui hoje você trabalhar de domingo a domingo praticamente pra ganhar um salário mínimo se matando... é ( ) e você vai fazer o quê? (D.S.cdt D.M.sdt P MM 01)

4. Situação de preservação de face positiva
Comando:

Maltratar animais é crime e prevê pena de 3 meses a um ano de detenção. Bruna e Letícia presenciaram o vizinho espancando um cachorro. Bruna pensou de imediato em acionar a polícia militar ambiental, já Letícia pensou em prestar atendimento ao animal.

Excerto 4:

F1: *é... questão de saúde a gente tá falando de saúde... de pessoas assim agora a questão sobre animais... que a gente sabe que muita gente... maltrata os animais né nas <<ru>> os animais que são... que estão nas ruas... sem abrigo e aí tem as vezes pessoas ou até mesmo pessoas que tem animais em casa e maltrata espanca... os animais né... e aí... tipo assim se você visse alguém maltratando um animal você imediatamente você ia ajudar... aquele animal assim resgatar aquele animal pra tentar ajudar levar pra um hospital ou você poderia já imediatamente já ligar pra polícia... ver o que tava acontecendo... como é que seria a sua reação?...*

F2: olhe eu eu não sei eu nun- nunca presenciei um caso desse assim eu não sei qual seria a minha reação se... eu eu acho que provavelmente seria de primeiro tentar ajudar o animal... eu não sou uma pessoa que gosta de ter animais em casa... mas também não gosto de ver... é... maltratá-los né... eu acho assim se você principalmente se for a pessoa que cria... pega um animal pra criar... a partir de um certo tempo por principalmente a gente vê acontecer casos de cachorro ou gato fica velho a pessoa já não quer mais... (...)

5. Situação de preservação de face negativa
Comando:

Luana e Brena estavam relembrando da primeira vez delas. Em meio a tantas gargalhadas, a mãe das duas chegou sem que elas percebessem e descobriu que elas perderam a virgindade muito antes do que ela imaginava, o que a deixou muito magoada.

Excerto 5:

F1: *é o seguinte agora essa aqui... é muita...*
F2: ((RISOS)) (o quê que eu digo?) quero responder esse não... passa <<pra>> outra...
F1: (hes) em relação à sua primeira vez... assim ficou sabendo você e o seu namorado no caso... mas... ã?...
[
F2: passe <<pra>> outra eu prefiro ( ) eu prefiro que você passe <<pra>> outra...
F1: *<<pá>> frente? <<pra>> outra... mas ( ) ser besteira... eu ia perguntar qual foi a primeira pessoa que você chegou se você chegou <<pra>> mãe <<pra>> conversar sobre isso...* (W.S.$_{cdt}$ A.G.$_{sdt}$ D M_F 28)

CONSIDERAÇÕES FINAIS

Esperamos ter contribuído com uma metodologia de coleta de dados, que se mostrou produtiva para captar nuanças de polidez. A amostra "Rede Social de Informantes Universitários" está disponível para toda a comunidade acadêmica, podendo subsidiar várias pesquisas que irão contribuir para a descrição do português falado no agreste sergipano, a partir de dados de fala de informantes da comunidade de prática, de alunos da Universidade Federal de Sergipe, *campus* Prof. Alberto Carvalho, situado no município de Itabaiana/SE, e consequentemente para a descrição do português falado no Brasil. Até o momento, esse banco de dados já subsidiou a pesquisa de Araujo (2014), sobre os efeitos da polidez no uso do futuro do pretérito em português e a de Santos (2014), sobre a variação de "nós/a gente" e a polidez.

REFERÊNCIAS

ARAUJO, A. S. *Você me faria um favor?: o futuro do pretérito e a expressão de polidez*. Sergipe, 2014. Dissertação (Mestrado em Letras) – Centro de Educação e Ciências Humanas, Universidade Federal de Sergipe.

BATTISTI, E. O estudo sociolinguístico da variação. In: *Anais do* CELSUL, 2008, p. 1-13.

BLAKE, R.; JOSEY, M. The /ay/ diphthong in Martha's Vineyard community: what can we say 40 years after Labov? *Language in Society*, Cambridge: Cambridge University Press, v.4, n. 32, p.451-485, 2003.

BROWN, P.; LEVINSON, S. C. *Politeness: some universals in language usage*. Cambridge: Cambridge University Press, 1987.

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

FREITAG, R. M. K. Banco de dados falares sergipanos. *Working Papers em Linguística*, Florianópolis: PPGLg, v. 14, p. 156-164, 2013.

_____. O controle dos efeitos estilísticos dos papéis sociopessoais e do sexo/gênero na entrevista sociolinguística. In: *Anais do* II *Congresso Internacional de Dialetologia e Sociolinguística* – CIDS, p. 289-296, 2012.

FREITAG, R. M. K.; MARTINS, M. A.; TAVARES, M. A. Bancos de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: potencialidades e limitações. *Alfa: Revista de Linguística*, São Paulo: v. 56, n.3, p. 917-944, 2012.

MEYERHOFF, M. *Introducing Sociolinguistics*. New York: Routledge, 2006.

MILROY, L. Social Networks. In: CHAMBERS, J. K.; TRUDGILL, P.; SCHILLING-ESTES, N. Eds. *The handbook of language variation and change*. Oxford Blackwell Publishing, 2002.

OUSHIRO, L. *Uma análise variacionista para as Interrogativas-Q*. São Paulo, 2011. Dissertação (Mestrado em Semiótica e Linguística). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

SANTOS, K. C. *Estratégias de polidez e a variação de nós vs. a gente na fala de discentes da Universidade Federal de Sergipe*. Sergipe, 2014. Dissertação (Mestrado em Letras). Centro de Educação e Ciências Humanas, Universidade Federal de Sergipe.

CAPÍTULO 9

TRANSCRIÇÃO DE ENTREVISTAS SOCIOLINGUÍSTICAS COM O ELAN

Livia Oushiro

INTRODUÇÃO

O ELAN (HELLWIG; GEERTS, 2013) é um programa para anotação de arquivos de áudio e vídeo, desenvolvido pelo Instituto Max Planck de Psicolinguística. Entre suas principais vantagens, encontram-se:

- a sincronização entre o arquivo de mídia e a transcrição/anotação, o que facilita enormemente a análise linguística dos dados (por exemplo, para codificação de variantes de variáveis fonéticas);
- a possibilidade de criação de múltiplas trilhas, que propicia não só a separação da fala de diferentes participantes, mas também a anotação detalhada de outros aspectos linguísticos e contextuais, bem como a representação de ações simultâneas (por exemplo, sobreposição de vozes, ações gestuais concomitantes às verbais);
- ferramentas mais sofisticadas de buscas dentro de um *corpus* (por exemplo, para encontrar todas as ocorrências de variantes de uma variável);
- a ampla flexibilidade de formatos de exportação da transcrição (.txt, .textgrid etc.) e, consequentemente, a compatibilidade com outros programas como Word, Excel, R, Rbrul, Praat, etc.;
- o fato de ser gratuito, e que vem sendo utilizado cada vez mais entre estudiosos da língua em uso, entre os quais os sociolinguistas (NAGY;

MEYERHOFF, 2013), o que configura um ponto positivo no compartilhamento de dados e metodologias.

O programa está disponível em <http://www.lat-mpi.eu/tools/elan/>. Nessa página, clique em *Download the latest version*. Na página que se abrir, baixe a versão correspondente a seu sistema operacional (Windows, Mac OS ou Linux). Além do programa, nesse site é possível baixar um manual detalhado e inscrever-se na lista de discussão. Após instalá-lo, você pode mudar a língua da interface clicando em *Opções* > *Língua* > *Português*. O principal objetivo deste tutorial é apresentar as ferramentas que serão de maior utilidade para transcrições de entrevistas sociolinguísticas. Ao final, apresentam-se algumas breves notas sobre a ferramenta de buscas, a exportação de arquivos e normas gerais de transcrição.

1. ARQUIVOS DE ANOTAÇÃO

Ao iniciar o ELAN, aparecerá uma janela vazia. Para começar uma nova anotação, clique em *Arquivo* > *Novo*..., ou use o atalho [Ctrl] + [N][1]. Uma nova janela se abrirá (Figura 1). Do lado esquerdo, selecione o arquivo de som que deseja transcrever e clique no botão [>>] para adicioná-lo aos *Arquivos Selecionados* à direita. É preferível utilizar os arquivos de som em formato .wav (não .mp3), pois eles permitem a visualização da onda sonora[2]. Clique em *OK*. O arquivo será aberto em uma nova janela.

1 No Mac, substitua a tecla [Ctrl] por [Command].

2 É possível que, mesmo abrindo o arquivo de áudio no formato .wav, a onda sonora não esteja tão claramente visível. Para melhorar a visualização, vale a pena ajustar a amplitude da onda. Isso pode ser feito no programa Audacity, disponível gratuitamente em <http://audacity.sourceforge.net/download/>. Após instalar o programa faça, primeiramente, uma cópia do arquivo sonoro que você deseja amplificar – é sempre bom manter uma cópia do original, sem alterações! Abra o novo arquivo no programa. No menu superior, clique em *Effect* > *Amplify*.... Na janela que abrir, aumente a amplitude através do slider. Clique em *Allow clipping* e clique em OK. No menu superior, clique em *File* > *Export*... e salve o arquivo sonoro com onda amplificada. Use esse novo arquivo para trabalhar no ELAN.

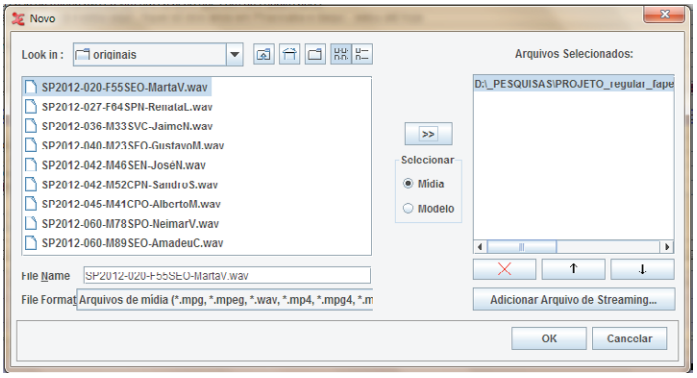


Figura 1 – Seleção de um arquivo de mídia para criação de nova anotação.

O ELAN organiza os arquivos em *projetos*. Cada projeto consiste em pelo menos dois arquivos: um ou mais *arquivos de mídia* (em geral, em estudos sociolinguísticos, um arquivo de som, mas também é possível fazer anotações de arquivos de vídeos), e um *arquivo de anotação* (com a extensão .eaf, se criado no ELAN, mas é possível inserir outros formatos de arquivo).

O ELAN associa um arquivo de anotação a seu respectivo arquivo de áudio, e armazena tais informações no arquivo .eaf. Isso significa que cada vez que você abre um arquivo do ELAN, o programa automaticamente busca um arquivo de som associado a ele. Se o ELAN não encontrar o arquivo sonoro no local especificado previamente (porque você está abrindo o arquivo de anotação em outro computador ou porque o moveu para um novo local no disco rígido), o programa pedirá novamente a localização desse arquivo (Figura 2).

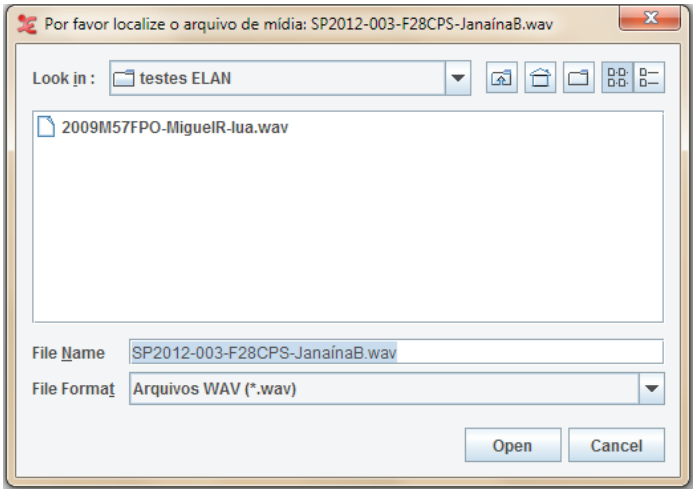


Figura 2 – Janela para localizar o arquivo de mídia.

Para abrir um arquivo de anotação criado previamente, clique em *Arquivo* > *Abrir*..., ou use o atalho [Ctrl] + [O]. Para salvar o seu trabalho, clique em *Arquivo* > *Salvar*..., ou o atalho [Ctrl] + [S].

2. AS PRINCIPAIS FUNÇÕES DO ELAN

Ao abrir um arquivo de áudio para transcrição, a janela principal do ELAN aparecerá (Figura 3). Ela contém diversas ferramentas que são descritas detalhadamente a seguir.

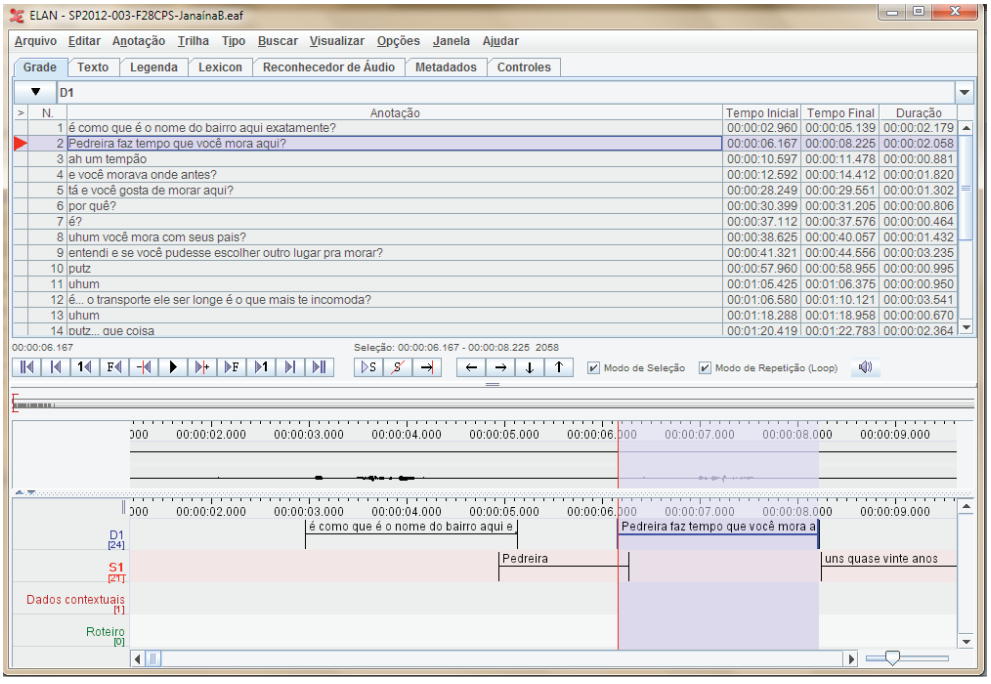


Figura 3 – Janela principal do ELAN.

Para nossos propósitos, as abas *Controles*, *Grade* e *Texto* na metade superior da janela principal do ELAN são as mais relevantes. A aba *Controles* tem dois *sliders*, que permitem ajustar o volume e a velocidade da gravação. Este último pode ser útil para transcrever falas muito rápidas (Figura 4).

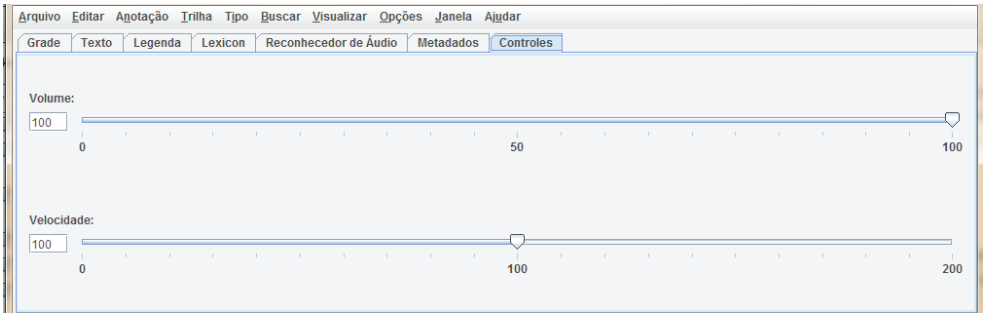


Figura 4 – Janela principal do ELAN – Controles de volume e velocidade.

A aba *Grade* apresenta uma tabela de unidades de anotação de uma determinada trilha, junto com as informações de tempo (inicial, final, duração) para cada anotação (Figura 5). Elas podem ser selecionadas (com um clique) e editadas (com duplo clique). A anotação selecionada é destacada com uma caixa azul escura e indicada por um triângulo vermelho. A Grade é sincronizada com a onda sonora, de modo que ao selecionar uma anotação, automaticamente a porção sonora é exibida no *Visualizador da Onda Sonora*. De modo inverso, a seleção de um trecho sonoro no visualizador destaca o intervalo correspondente na Grade.

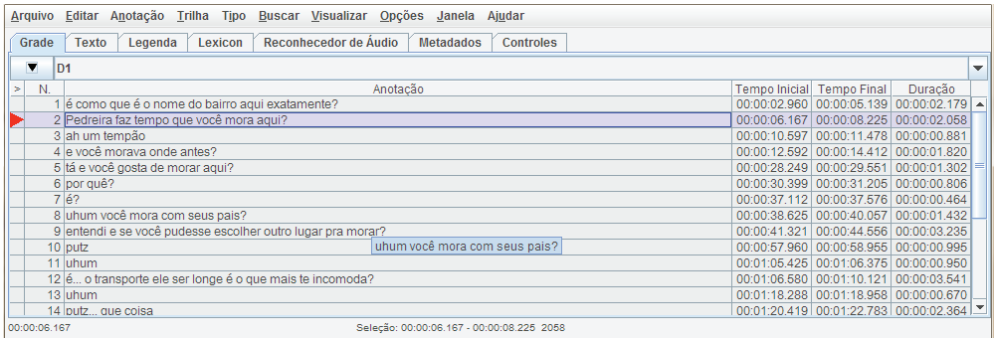


Figura 5 – Janela principal do ELAN – Grade.

A aba *Texto* apresenta um texto corrido de todas as transcrições em uma determinada trilha (Figura 6). As fronteiras entre as unidades de anotação são indicadas por pontos. De modo semelhante à aba *Grade*, uma unidade de anotação pode ser selecionada com um clique e editada com o duplo-clique (através de uma janela de edição na qual a transcrição pode ser modificada), e a onda sonora e o texto são sincronizados.

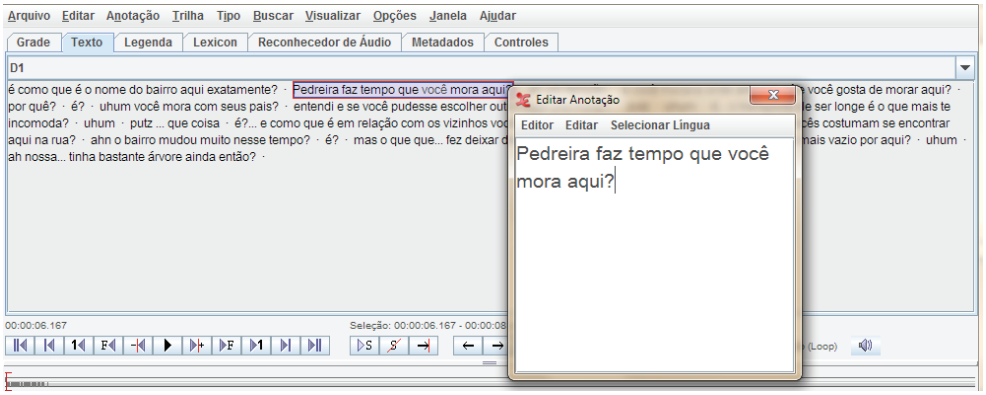


Figura 6 – Janela principal do ELAN – Texto.

Na parte central da janela, os *Botões de Controle* permitem tocar ou pausar o arquivo de áudio, navegar pela gravação, selecionar partes do arquivo de áudio e navegar entre as anotações (Figura 7).

Figura 7 – Janela principal do ELAN – Botões de controle.

O primeiro grupo de botões permite tocar ou pausar a gravação, e navegar pelo arquivo de áudio. Da esquerda para a direita, as respectivas funções são: (i) ir para o começo da mídia; (ii) ir para o enquadramento anterior; (iii) voltar um segundo; (iv) voltar um *frame*; (v) voltar um *pixel*; (vi) tocar/pausar; (vii) ir ao próximo *pixel*; (viii) ir ao próximo *frame*; (ix) adiantar um segundo; (x) ir para o próximo enquadramento e (xi) ir para o final da mídia. O segundo conjunto de botões controla as seleções. Da esquerda para a direita: (i) tocar o intervalo selecionado; (ii) limpar o intervalo selecionado e (iii) mover a linha vermelha para os extremos da seleção. O terceiro conjunto de botões permite navegar para frente e para trás entre as unidades de anotação nas trilhas. Da esquerda para a direita: (i) ir para a anotação anterior; (ii) ir para a próxima anotação; (iii) ir para a trilha de cima; e (iv) ir para a trilha abaixo. Por fim, há duas caixas: (i) se *Modo de Seleção* estiver selecionada, um trecho tocado será automaticamente selecionado; (ii) se *Modo de Repetição (Loop)* estiver selecionada, o intervalo selecionado será tocado repetidas vezes ao clicar sobre o botão *Play*.

O *Visualizador de Densidade de Anotações* se localiza abaixo dos botões de controle (Figura 8) e fornece uma rápida impressão visual de quanto do arquivo sonoro já foi transcrito. As marcações em cinza indicam as regiões no áudio que

contêm unidades de anotação, de modo que você pode visualizar rapidamente quais partes das gravações já foram transcritas. Esse visualizador também é um modo fácil de navegar pelo arquivo, já que a extensão da barra corresponde ao da gravação completa, independente do *zoom* no *Visualizador da Onda Sonora*. A barrinha vertical vermelha indica a posição atual do cursor. É possível navegar facilmente pelo arquivo ao mudar a barrinha para a esquerda ou para a direita.

Figura 8 – Janela principal do ELAN – Visualizador de densidade de anotações.

Na metade inferior da janela principal do ELAN, o *Visualizador da Onda Sonora* (Figura 9) apresenta a amplitude (eixo vertical) ao longo do tempo (eixo horizontal) do arquivo de áudio. Ele também mostra a atual posição do cursor (linha vertical vermelha), informação de tempo e quais partes do arquivo de áudio estão selecionadas (destacadas em azul claro). Você pode aumentar ou diminuir a escala de tempo visualizada segurando o botão [Ctrl] e usando a roda de rolagem de seu *mouse*.

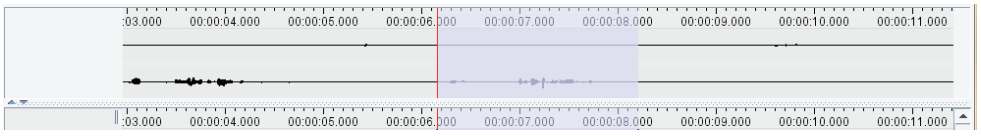


Figura 9 – Janela principal do ELAN – Visualizador da Onda Sonora.

Por fim, abaixo do *Visualizador da Onda Sonora* estão as *Trilhas de Anotação*, sendo uma trilha por falante. Cada trilha contém unidades de anotação, que por sua vez contêm a transcrição. No lado esquerdo do painel estão os nomes das trilhas; aquela selecionada ou ativa está em vermelho claro (por exemplo, na Figura 10, a trilha selecionada é S1). Ao posicionar o cursor do *mouse* sobre os nomes (sem clicar), aparecerá uma janela com informações mais específicas sobre a trilha, com o seu nome, nome do participante, nome do anotador, etc.

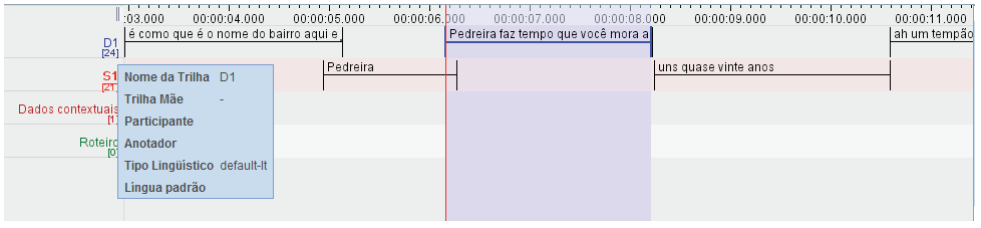


Figura 10 – Janela principal do ELAN – Trilhas de anotação.

3. TRILHAS

De modo geral, uma trilha de anotação no ELAN contém um texto transcrito junto com marcações de tempo. Essas marcações são usadas pelo programa para sincronizar a exibição da onda sonora e da anotação.

As trilhas no ELAN podem ter diferentes funções. Por exemplo, para um mesmo falante, você pode criar uma trilha para a transcrição ortográfica das sentenças, outra para as palavras individuais, outra para as unidades morfológicas, outra para unidades fonéticas, etc. Para estudos sociolinguísticos, é interessante criar pelo menos as seguintes trilhas:

- uma trilha para cada participante da gravação (documentador, informante, falantes adicionais, etc.);
- uma trilha para dados contextuais, que se referem à ações simultâneas dos participantes ou de outros eventos no ambiente de gravação, mas não aos enunciados da entrevista [risos], [tosse], [barulho de moto], etc.;
- uma trilha para identificar partes do roteiro (bairro, infância, lista de palavras etc.)

Para criar uma nova trilha, clique em *Trilha* > *Adicionar nova trilha*... ou [Ctrl] + [T]. Uma nova janela se abrirá, com uma lista na metade superior e diferentes abas e campos na metade inferior (Figura 11). A aba *Adicionar* estará destacada. Nessa janela, insira as seguintes informações para cada trilha que você deseja criar:

- nome da trilha: inserir a identificação do falante (por exemplo, "D1" para documentador, "S1" para o informante, "S2, S3" para falantes adicionais) ou de outros tipos de anotação (por exemplo, "Dados Contextuais", "Roteiro");
- participante: inserir o pseudônimo dos informantes e o nome dos documentadores;
- anotador: inserir o seu nome completo.

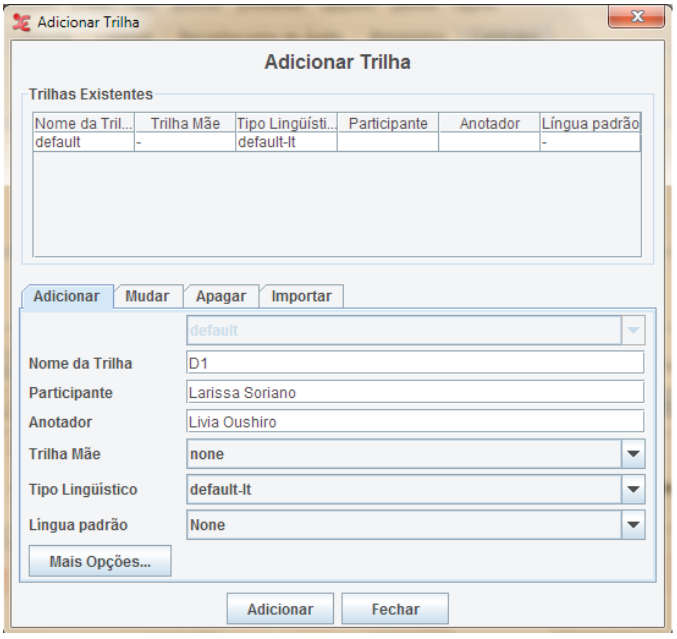


Figura 11 – Janela para criação de novas trilhas.

Os demais campos ("Trilha mãe", "Tipo linguístico", "Língua padrão") podem ser deixados nos valores pré-configurados. Para cada trilha, clique em "Adicionar". Depois disso, clique sobre a aba "Apagar" para deletar a trilha "default". Por fim, clique em "Fechar" para sair da janela.

4. PROCEDIMENTOS

4.1. Mudança nos atalhos

Em geral, é preferível trabalhar com o teclado o máximo possível, pois o trabalho se torna muito mais rápido e menos cansativo (fisicamente) do que alternar entre o *mouse* e o teclado. Para isso, você pode usar os atalhos para a maior parte dos comandos no ELAN. Para ver (e imprimir, se quiser) uma lista de atalhos, clique em *Visualizar* > *Atalhos*....

O ELAN vem com um grande conjunto de atalhos pré-configurados, alguns dos quais não muito intuitivos. Você pode mudar quaisquer atalhos para uma combinação que achar mais adequada, através de *Editar* > *Preferências* > *Editar Atalhos*.... Uma nova janela se abrirá (Figura 12). Selecione *Categoria* do menu, no canto inferior esquerdo, para organizar os diferentes atalhos de acordo com sua função. Selecione o

atalho a editar clicando sobre ele e clique em *Editar Atalho* (ver Figura 13). Insira a combinação de teclas que você prefere na nova janela e clique em *Apply in all modes*. Após fazer todas as alterações que desejar, clique no botão *Salvar* na janela de atalhos.

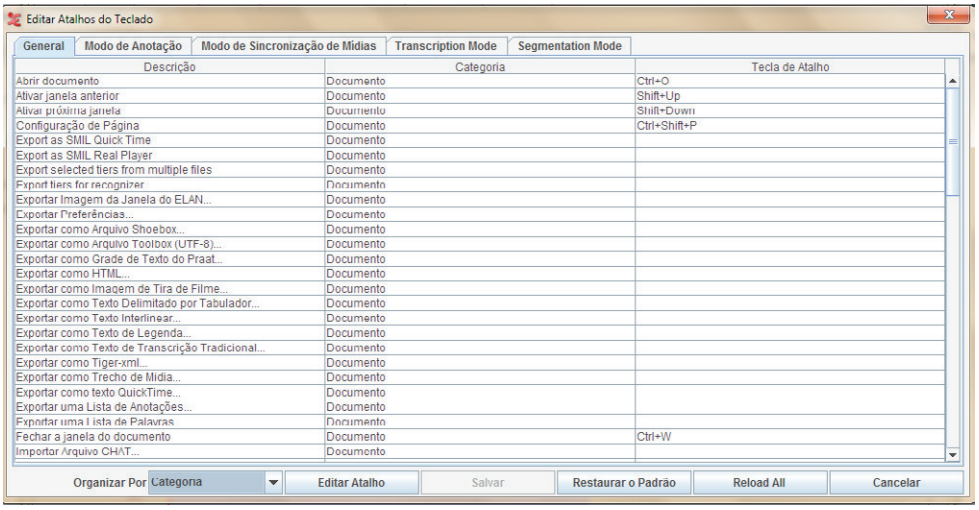


Figura 12 – Janela de edição de atalhos.

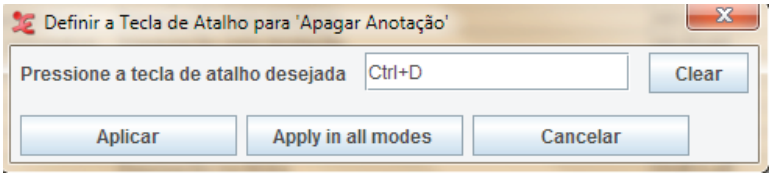


Figura 13 – Especificação de novo atalho.

4.2. Sugestões de atalhos

As seguintes teclas de atalho são sugeridas para facilitar a transcrição no ELAN.

Modo de anotação – Edição de anotações	
Deletar anotação	[Ctrl] + [Delete]
Modificar valor da anotação	[Ctrl] + [M]
Modificar o tempo da anotação	[Ctrl] + [Enter]
Nova anotação aqui	[Shift] + [Enter]
Remover valor da anotação	[Shift] + [Delete]
Modo de anotação – Trilha	
Ativar trilha superior	[Ctrl] + [↑]
Ativar trilha inferior	[Ctrl] + [↓]
***Transcription Mode* – Navegação pela anotação**	
Ir para a próxima anotação	[Alt/Option] + [→]
Ir para a anotação anterior	[Alt/Option] + [←]
***Segmentation Mode* – Navegação na mídia**	
Tocar/pausar a mídia	[Shift] + [Space]
Tocar seleção	[Ctrl] + [Space]
Definir tempo 1 segundo atrás	[Ctrl] + [←]
Definir tempo 1 segundo à frente	[Ctrl] + [→]
Ir para o *pixel* anterior	[Shift] + [←]
Ir para o próximo *pixel*	[Shift] + [→]
***Segmentation Mode* – Seleção**	
Limpar seleção	[Esc]
Mover cursor para os extremos da seleção	[Ctrl] + [/]

Quadro 1 – Sugestão de novos atalhos no ELAN.

Também é recomendável mudar as preferências de edição, de maneira que para salvar as mudanças em uma caixa de anotação, baste pressionar [Enter] (a configuração *default* é pressionar [Ctrl] + [Enter]). Para fazer isso, clique em *Preferências* > *Editar Preferências*.... No menu à esquerda, clique em *Editar* e, à direita, selecione a segunda opção: *A tecla Enter realiza as mudanças na caixa de edição alinhada*. Você também pode escolher a primeira opção: *Ao retirar a seleção da caixa de edição do texto alinhado as mudanças serão realizadas*, o que significa que, quando você sair da caixinha de edição (por exemplo, clicando fora

dela), o que foi digitado dentro da caixinha será salvo (a configuração *default* é que o ELAN descarte tais mudanças).

4.3. Fluxo de trabalho

Os seguintes passos são sugeridos para um fluxo de trabalho mais dinâmico na transcrição das gravações:

1) ligue o *Modo de Seleção* (e o *Modo de Repetição*, se quiser) clicando sobre as respectivas caixas de seleção acima da onda sonora;
2) o cursor estará no início do arquivo. Pressione [Shift] + [Space] para começar a tocar. À medida que o cursor se mover, ele selecionará o intervalo tocado;
3) deixe o cursor tocar até antes da primeira sentença. Pressione [Shift] + [Space] para pausar a gravação;
4) use os controles de navegação de mídia para mover o cursor exatamente para o ponto em que você quer começar um nova anotação. As teclas [Ctrl] + [←] e [Ctrl] + [→] voltam/adiantam 1 segundo e as teclas [Shift] + [←] e [Shift] + [→] voltam/adiantam um *frame* (provavelmente esses últimos serão mais utilizados);
5) desfaça a seleção atual com [Esc].
6) inicie o *playback* novamente com [Shift] + [Space]. O cursor agora vai começar a selecionar o trecho de fala.
7) pause o *playback* logo após o final da fala com [Shift] + [Space]. Se necessário, use os controles de navegação de mídia novamente para mover o cursor exatamente para o ponto final da nova anotação. Você pode ouvir o trecho selecionado com os comandos [Ctrl] + [Space];
8) ative a trilha do falante correspondente onde você quer a nova anotação usando [Ctrl] + [↑]/[↓];
9) se você ligou o *Loop Mode*, pressione [Ctrl] + [Space] para começar a tocar a seleção;
10) pressione [Shift] + [Enter] para criar uma nova anotação na trilha ativa;
11) após transcrever a fala, pressione [Enter] para salvar a transcrição;
12) pressione [Ctrl] + [Space] para parar o *loop* da seleção atual;
13) pressione [Shift] + [Space] para recomeçar o *playback* da atual posição do cursor;
14) repita os passos (3) a (13).

Com um pouco de prática, o usuário perceberá que tais comandos logo se tornam automatizados. Para *Editar o conteúdo de uma anotação*, navegue para a anotação com [Alt/Option] + [←]/[→]. Isso vai mover a anotação selecionada para trás ou para frente. Pressione [Ctrl] + [M]. Uma janela de edição se abrirá. Após fazer as modificações desejadas, pressione [Enter] para salvar e fechar a janela de edição. Para *Mudar o tempo de duração* dos segmentos, certifique-se primeiro de que o *Modo de Seleção* está ligado. Navegue até a anotação com [Alt/Option] + [←]/[→]. Use [Ctrl] + [/] para colocar o cursor do lado da seleção que você deseja mudar (esquerda ou direita). Use os comandos de navegação de mídia para mudar a extensão da seleção ([Ctrl] + [←]/[→] move 1 segundo, [Shift] + [←]/[→] move 1 *pixel*). Após ajustar a seleção, pressione [Ctrl] + [Enter]. Para *apagar uma anotação*, navegue até ela [Alt/Option] + [←]/[→] e pressione [Ctrl] + [Delete].

5. BUSCAS EM MÚLTIPLOS ARQUIVOS E EXPRESSÕES REGULARES

Uma vez que o *corpus* esteja transcrito, é possível buscar certas palavras ou sequências de caracteres em um ou mais arquivos de transcrição no próprio ELAN. Essas opções se encontram em *Buscar*[3]. Uma das vantagens do programa é permitir a busca não só por sequências literais (por exemplo, "os", "as"), mas também através de *expressões regulares*, que são sequências de caracteres em que alguns têm significado especial. Pode-se especificar, por exemplo, que se deseja buscar todas as ocorrências de /r/ em coda em posição final com a expressão regular [aáâeéêiíoóôuú]r\b, em que "[aáâeéêiíoóôuú]" cria uma classe de caracteres que podem se alternar (neste caso, qualquer vogal oral, especificadas por seus grafemas) e "\b" indica uma fronteira de palavra.

Para utilizar expressões regulares, o usuário deve selecionar a caixa *Expressão regular*, para indicar ao programa que não se deseja buscar a sequência literal. Uma lista completa de caracteres especiais no ELAN pode ser consultada em seu manual (HELLWIG; GEERTS, 2013, p. 305).

3 Para múltiplos arquivos, deve-se definir um *Domínio de busca*, que se refere a uma pasta ou a um conjunto de arquivos em que a busca deve ser realizada.

6. EXPORTAÇÃO DE ARQUIVOS

Como visto, os arquivos do ELAN são salvos com extensão .eaf. O programa permite a exportação desses arquivos em diversos formatos (texto delimitado por tabulador, texto de transcrição tradicional, lista de palavras, textgrid do Praat, etc.), compatíveis com outros programas como Word (.txt), Excel (texto delimitado por tabulador), R (.txt) e Praat (.textgrid). Tais opções se encontram em *Arquivo* > *Exportar como*.... Recomenda-se que o usuário explore as diversas opções de exportação de arquivo, para que encontre aquela que melhor atenda aos seus propósitos.

7. ALGUMAS NOTAS SOBRE NORMAS DE TRANSCRIÇÃO

O objetivo principal de um sistema de transcrição é transpor a língua falada para o texto escrito de uma forma fiel à língua oral, mas inteligível, de modo a armazenar o material coletado em meio escrito de maneira padronizada para facilitar a sua análise. É necessário lembrar que nenhum sistema de transcrição é capaz de reproduzir a fala tal e qual e que qualquer sistema proposto sofrerá de limitações; com esse fato em mente, os critérios devem sempre levar em conta os objetivos do grupo de pesquisa e os tipos de análise que serão desenvolvidas com o material. Compare, por exemplo, diferentes normas propostas para o Projeto da Norma Urbana Oral Culta (NURC) (CASTILHO; PRETTI, 1986); o projeto Amostra Linguística do Interior Paulista (ALIP) (TENANI; GONÇALVES, s/d); o C-ORAL-BRASIL (MELLO; RASO, 2009); o Projeto SP2010 (MENDES; OUSHIRO, 2013).

Entretanto, é possível traçar algumas recomendações gerais. Quanto maior o número de critérios, mais heterogêneas tendem a ser as transcrições das gravações do mesmo *corpus*, algo que não é desejável quando um dos principais objetivos é padronizar o material coletado. O fato de que o pesquisador terá a transcrição sincronizada com a onda sonora leva à preferência por um sistema mais simples de transcrição – por exemplo, sem o uso (ou abuso) de caracteres como dois pontos, chaves, parênteses, maiúsculas e minúsculas, a anotação de dados contextuais no meio da fala dos informantes e a indicação de apagamento de segmentos, por exemplo, "pe(i)xe", "fala(r)". Muitos pesquisadores transcrevem suas gravações de acordo com a variável a ser analisada, adotando diferentes convenções para cada variante. No entanto, é interessante projetar que tais gravações poderão ser utilizadas em pesquisas futuras, pelo mesmo pesquisador ou por outros, de modo que se recomenda uma transcrição mais "neutra".

É importante transcrever por extenso tudo o que se ouviu, sobretudo números (por exemplo, "terceiro" em vez de "3°"), pois tais palavras podem conter segmentos-alvo para determinados estudos. Siglas e abreviaturas podem ser grafadas de formas diferentes a depender se são pronunciadas letra a letra (por exemplo, "B.O.", "I.N.S.S.", "P.T.") ou como palavras (por exemplo, "USP", "IAMSP", "UFSCAR"). Podem-se criar convenções para marcação de pausas, alongamentos, truncamentos, hesitações e interjeições; no entanto, é preferível que tais anotações não conduzam à criação de novas palavras (por exemplo, indicar silabação por hífen: "com-ple-ta-men-te"), para que não se "inflacione" o número total de palavras do *corpus*. Se se pretende compartilhar o *corpus* com outros pesquisadores, vale a pena revisar as transcrições, preferivelmente por outra pessoa que não o próprio transcritor (para fugir dos vícios de um texto já familiar).

Por fim, a regra mais básica é a aplicação consistente de qualquer norma criada; o pesquisador poderá, então, trabalhar com maior confiança em suas futuras explorações da amostra transcrita.

REFERÊNCIAS

CASTILHO, A.; PRETI, D..*A linguagem falada culta na cidade de São Paulo*: materiais para seu estudo, vol. I – Elocuções Formais. São Paulo: T.A. Queiroz, 1986.

HELLWIG, B.; GEERTS, J. ELAN – Linguistic Annotator. Versão 4.4.0. Disponível em: <http://www.mpi.nl/corpus/manuals/manual-elan.pdf>. Acessado em: 11 fev. 2014.

MELLO, H.; RASO, T. Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. *Revista Portuguesa de Humanidades*, Portugal, v.13, n. 1, p. 301-325, 2009.

MENDES, R. B.; OUSHIRO, L. Documentação do Projeto SP2010 – Construção de uma amostra da fala paulistana. Universidade de São Paulo, 2013. Disponível em: <http://projetosp2010.fflch.usp.br/producao-bibliografica>. Acessado em: 11 fev. 2014.

NAGY, N.; MEYERHOFF, M. Extending ELAN into Variationist Sociolinguistics. Workshop apresentado no NWAV42, Pittsburgh-PA/EUA, 2013. Disponível em: <http://www.nwav42.pitt.edu/workshops/#elan>. Acessado em: 11 fev. 2014.

TENANI, L. E.; GONÇALVES, S. C. L. *Manual do Sistema de Transcrição de Dados* - Projeto ALiRP (Amostra Linguística de Rio Preto). Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, s/d.

CAPÍTULO 10

TRATAMENTO DE DADOS COM O R PARA ANÁLISES SOCIOLINGUÍSTICAS

Livia Oushiro

INTRODUÇÃO

Ao se debruçar sobre a língua em uso, o linguista inevitavelmente se depara com a variação linguística. No português falado no Brasil, encontram-se, por exemplo, diversas realizações para o "r", tanto em posição de ataque quanto em coda silábica (como tepe, vibrante múltipla, retroflexo, fricativa velar, etc.); o emprego dos pronomes "tu", "você", "ocê", "cê", para se referir ao interlocutor; o uso de sintagmas verbais plurais ora com marcação de número apenas no sujeito (por exemplo, "eles foi"), ora com marcação redundante no sujeito e no verbo (por exemplo, "eles foram").

Um dos principais fundamentos dos estudos sociolinguísticos é a premissa de que a variação linguística – verificada em todas as línguas, em todas as comunidades e, em última instância, na fala de um mesmo indivíduo – faz parte do sistema linguístico e da competência comunicativa dos falantes. A variação linguística não só é inerente, como também é ordenada (WEINREICH; LABOV; HERZOG, 1968): as flutuações observadas formam padrões que podem ser descritos e analisados pelo estudioso da língua em uso.

A observação desses padrões, no entanto, requer a análise de uma grande quantidade de dados. A partir da observação de poucas ocorrências, de um ou poucos falantes, dificilmente se poderia chegar a conclusões confiáveis sobre quais falantes tendem a empregar uma ou outra forma, em quais contextos (linguísticos

ou sociais) elas tendem a ocorrer e por que a variação ocorre do modo como se observa. A sociolinguística variacionista se assenta sobre o Paradigma Quantitativo (BAYLEY, 2002; GUY, 1993), que busca modelar a competência comunicativa dos falantes através da análise de formas linguísticas variáveis em seus contextos de uso, a fim de derivar afirmações acerca da probabilidade de coocorrência de uma forma linguística variável e as características contextuais.

Desse modo, o sociolinguista variacionista lida com uma grande quantidade de dados. Entre as suas diversas tarefas, incluem-se: (i) a coleta de dados (em geral, na forma de gravações de entrevistas sociolinguísticas com falantes de uma comunidade); (ii) a transcrição dessas gravações; (iii) a definição de uma variável sociolinguística e de seus contextos linguísticos possíveis (o contexto variável); (iv) a identificação de ocorrências no *corpus* de entrevistas; (v) o levantamento de hipóteses sobre fatores, de natureza social e linguística, que estejam correlacionados ao uso da variável; (vi) a codificação das ocorrências de acordo com as hipóteses levantadas; (vii) a análise quantitativa dos dados no GoldVarb X ou RBrul e (viii) a interpretação de resultados obtidos.

Algumas dessas tarefas exigem conhecimento especializado e criatividade para ser bem executadas – por exemplo, a coleta de boas gravações, o levantamento de hipóteses, a interpretação de resultados. Algumas outras podem ser bastante repetitivas, mecânicas e previsíveis – como a identificação de ocorrências no *corpus* (quando já se definiu a variável e seu envelope de variação) e a sua extração para codificação. Para esse segundo conjunto de tarefas, em princípio, não é necessário um conhecimento especializado; por exemplo, não é necessário saber o que é um fonema para copiar e colar certos trechos de texto em uma planilha de codificação.

As tarefas repetitivas, mecânicas e previsíveis podem ser automatizadas através do uso do computador. Nesse sentido, o programa R (R CORE TEAM 2013) é de grande valia para a otimização do tempo empregado na execução dessas tarefas. O R é uma linguagem de programação voltada à análise de dados, que pode ser utilizada para realizar computações estatísticas e gráficas, compilar e anotar *corpora*, produzir listas de frequências, entre diversas outras tarefas. Uma de suas principais vantagens é o fato de ser gratuito e estar disponível para uma variedade de plataformas (UNIX, Windows e MacOS).

Sendo uma linguagem de programação, o R permite que o usuário customize uma série de tarefas que deseja executar e, consequentemente, tenha maior controle sobre os resultados obtidos. Isso significa, no entanto, que ao invés de clicar em botões com funções limitadas e pré-definidas, o usuário normalmente define as funções que deseja executar através de *linhas de comando*, que instruem o programa sobre o que fazer. Uma sequência de linhas de comando é chamada de *script* ou *código*. O exemplo (1) a seguir mostra um pequeno *script*, que instrui o

R a carregar um arquivo de transcrição na memória, apagar as marcas de parênteses e salvar o arquivo limpo.

(1)

```
> FabianaB<-scan(file=choose.files(),what="char",sep="\n")¶
> FabianaB.limpo<-gsub("\\(|\\)","",FabianaB)¶
> cat(FabianaB.limpo,file="FabianaB-limpo.txt",sep="\n")¶
```

Embora isso possa parecer complicado inicialmente, um pouco de prática levará o usuário a se familiarizar com o ambiente. Em geral, o esforço de criar um script só precisa ser feito uma vez, já que podemos salvar o código e reutilizá-lo quantas vezes forem necessárias, modificando apenas pequenas partes para readaptá-lo às novas demandas. Além disso, há uma série de scripts escritos previamente por outros usuários, na forma de *funções* e *pacotes*, que podemos baixar da internet e utilizar em nossas próprias tarefas. Tal é o caso das funções `identificacao()`, `extracao()` e `amostragem()`, do pacote `dmsocio`, que serão mais detalhadamente exploradas na seção 4 adiante. Antes de descrever a aplicação dessas funções, é necessário tratar da instalação do programa (seção 1), de conceitos básicos e algumas funções úteis para sua utilização (seção 2), e de arranjos prévios na organização de nosso *corpus* (seção 3). O artigo se encerra com uma visão perspectiva dessas funções no âmbito das análises sociolinguísticas e com a indicação de leituras adicionais para um maior aprofundamento das aplicações do R aos estudos linguísticos.

Não é demais salientar que, em se tratando de um tutorial prático, este texto foi pensado para ser lido com um computador à mão, de modo que o leitor possa reproduzir os exemplos durante a leitura. Não há como aprender a usar o R sem utilizá-lo. Portanto, mãos à massa!

1. INSTALAÇÃO DO PROGRAMA R

O primeiro passo para começar a utilizar o R é sua instalação. O programa pode ser baixado gratuitamente do site do Projeto R, no endereço <http://cran.r-project.org/>. Na seção *Download and Install R*, clique no link referente ao seu sistema operacional (Linux, MacOS ou Windows) e instale o R no diretório sugerido.

Inicie o programa. Você verá que o R possui uma interface bastante simples: um menu na parte superior com algumas opções usuais (Arquivo, Editar, etc.); alguns botões de comandos mais frequentes (Abrir script, Salvar, Imprimir, etc.); e uma janela do *Console* que informa a versão instalada e algumas breves notas

sobre o R. Na última linha, os sinais em (2) indicam que o R está pronto para receber comandos.

(2)

```
>|
```

Existe atualmente uma interface mais "amigável", chamada RStudio, que disponibiliza algumas ferramentas adicionais diretamente na interface gráfica, como a visualização dos *scripts* abertos recentemente, o histórico de linhas de comando executadas e a lista de pacotes instalados. O RStudio é, de fato, apenas uma interface gráfica e alternativa e sua utilização requer a instalação do programa R em seu computador (como fizemos na etapa acima). A instalação do RStudio é opcional, mas altamente recomendável. O programa pode ser baixado a partir do endereço <http://www.rstudio.com/ide/download/>. Ao escolher *Download RStudio Desktop*, o site identifica automaticamente o seu sistema operacional e mostra a versão adequada no item *Recommended for your system*. Baixe essa versão para o seu computador, instale e inicie o programa.

A interface do RStudio apresenta quatro janelas (Figura 1):

i. *Source*, para visualização e edição de *scripts*;
ii. *Environment* e *History*, com os objetos carregados na memória do R para a presente sessão e com o histórico de linhas de comando executadas;
iii. *Console*, no qual as funções e os *scripts* são executados;
iv. *Files*, *Plots*, *Packages*, *Help* e *Viewer*, respectivamente, para arquivos, gráficos, pacotes, ajuda e visualizador.

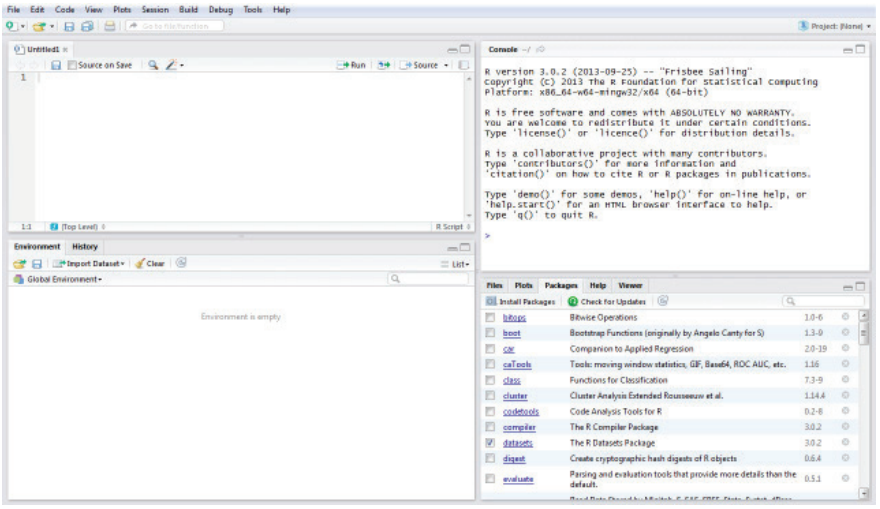


Figura 1 – Interface do RStudio.

Além de instalar o programa R e o RStudio, baixe os materiais de apoio a esse tutorial que estão disponíveis em <http://projetosp2010.fflch.usp.br/dmsocio>. Nessa página, clique sobre o *link Scripts+Exemplos+Minicorpora* e salve o arquivo em seu computador. Essa é uma pasta zipada que contém : (i) o arquivo "dmsocio.R", com o código para as funções `identificacao()`, `extracao()` e `amostragem()`; (ii) o arquivo "dmsocio-exemplos.R", com instruções simplificadas, modelos de como usar as funções do pacote dmsocio e os exemplos deste tutorial; (iii) os arquivos CodifR-12ent.txt e Ex26-MauricioB.txt; e (iv) dois *Minicorpora* (dmsocio-SP2010 e dmsocio-codif-R), que contêm cada qual a transcrição de 12 entrevistas sociolinguísticas do Projeto SP2010 (MENDES; OUSHIRO, 2013) e que serão utilizados na exemplificação da aplicação das funções. Há também uma pasta chamada "UTF-8", com os mesmos arquivos .txt em formato UTF-8 de codificação. Para abrir o arquivo zipado é necessário ter um descompactador, como os programas Winzip ou Zipeg. Depois de descompactá-lo, abra o arquivo "dmsocio-exemplos.R" no RStudio (*File* > *Open file*...), que usaremos na seção 4.

2. CONCEITOS BÁSICOS PARA USO DO PROGRAMA R

Como visto em (2) acima, os símbolos >| indicam que o programa R está pronto para receber comandos. Você pode, por exemplo, pedir que o R calcule o resultado de 2 + 3 como na linha de comando abaixo:[1]

(3)

```
> 2 + 3 ¶
```

Ao digitar 2 + 3 e pressionar ENTER, o R fornecerá a resposta da operação matemática dada em (3):

(4)

```
[1] 5
```

O número [1] simplesmente dá uma referência de quantos resultados existem para o cálculo requerido (para outras funções, o número de resultados pode estar na casa de centenas, milhares, milhões...). A resposta para a linha de comando em (3) é fornecida logo em seguida: 5. No entanto, deve estar claro que nosso verdadeiro interesse em utilizar o R não é como uma simples calculadora, mas que o programa realize operações muito mais complexas. Ainda em um exemplo simples, pode-se pedir que o R guarde o resultado dentro de um objeto nomeado pelo próprio usuário, para ser acionado posteriormente. Para isso, basta especificar o nome do objeto, seguido dos símbolos <-, que iconicamente indicam onde o R deve guardar tal informação. Desse modo, em vez do exemplo (3), podemos instruir que o R calcule o resultado de 2 + 3 e guarde-o em um objeto chamado **x** (ou qualquer outro nome que você queira dar[2]) do seguinte modo:

(5)

```
> x <– 2 + 3 ¶
```

1 Nos exemplos fornecidos daqui em diante, ">" indica que o R está disponível para novos comandos, e "¶" representa a tecla [ENTER], para que a linha de comando seja executada. Não se trata, portanto, de caracteres que devam ser digitados. Todos os exemplos deste artigo se encontram ao final do arquivo "dmsocio-exemplos.R". No RStudio, linhas de comando em um script aberto em *Source* são rodadas no console através de [CTRL] + [ENTER] (Windows) ou [Command] + [ENTER] (MacOS).

2 Em princípio, pode-se atribuir qualquer nome a um objeto. No entanto, a utilização de um mesmo nome para objetos diferentes fará com que o programa substitua o antigo.

Repare que, desta vez, o R não fornece o resultado do cálculo como fez em (4); ele pode ser acessado instruindo o R a mostrar qual é o conteúdo de **x** (ou outro nome que você tenha atribuído ao objeto), ao que o R responde 5:

(6)

```
> x ¶
[1] 5
```

Isso significa que **x** agora funciona como uma variável que equivale a 5 e pode ser utilizado em outras computações. Assim, se digitarmos a linha de comando **x+10**, o R retornará o resultado 15.

(7)

```
> x + 10 ¶
[1] 15
```

O objeto criado **x** deixará de ter o valor 5 até que a sessão corrente do R seja encerrada, até que o nome **x** seja atribuído a outro objeto (8), ou até que seu valor seja removido com a função `rm()` – *remover* (9):

(8)

```
x <- (16-4)*2 ¶
```

(9)

```
rm(x) ¶
```

Além de guardar valores numéricos, o R também pode guardar valores textuais, como caracteres, palavras, sentenças, transcrições de entrevistas sociolinguísticas e tabelas de dados, que concernem mais propriamente aos nossos interesses presentes. Tais informações podem ser armazenadas em diferentes tipos de objetos, chamados vetores e *dataframes*.

2.1. Vetores e dataframes

O tipo mais simples de uma estrutura de dados (e aquele que mais utilizaremos nos *scripts* para análises sociolinguísticas) é o *vetor*, uma sequência unidimensional de elementos como números ou caracteres (palavras, sentenças, textos). De fato, o objeto **x** criado acima é um vetor que contém apenas um elemento, 5.

Em lugar de valores numéricos, um vetor pode abarcar valores textuais, como na criação do vetor **y** abaixo:

(10)

```
> y <- "gato" ¶
```

Note que, para ser entendido como uma sequência de caracteres, o(s) valor(es) atribuído(s) ao vetor deve(m) vir entre aspas. Na próxima subseção, veremos como criar vetores com mais de um elemento.

O *dataframe* é uma estrutura de dados bidimensional, com linhas e colunas, equivalente a tabelas (como, por exemplo, uma planilha de dados do Excel). Colocado de outro modo, o *dataframe* é um conjunto de vetores de mesma extensão (mesmo número de linhas ou número de colunas). Para nossos propósitos, o *dataframe* será importante para a criação de planilhas de dados extraídos das transcrições de entrevistas sociolinguísticas, ou para que o R possa "ler" planilhas de dados previamente criadas.

2.2. Funções e argumentos

Como visto, o R funciona através de uma interface em que o usuário digita certas linhas de comandos a serem executados e o programa retorna os resultados no console ou os armazena em objetos. Na maior parte dos casos, as linhas de comando contêm *funções*, que instruem o programa sobre o que deve ser feito. Cada função contém um ou mais *argumentos*, que são variáveis pré-definidas e indicam, entre outras coisas, os dados com que se deseja trabalhar e como eles devem ser tratados.

O item (11) mostra a sintaxe básica de funções no R: ela consiste no nome da função e a especificação dos argumentos dentro de parênteses e separados por vírgulas.

(11)

```
funcao(arg. 1, arg. 2, arg. 3...)
```

Três funções serão empregadas na utilização dos *scripts* para análise de dados sociolinguísticos: `c()`, `getwd()/ setwd()`, e `read.table()`.

`c()`: Função genérica para combinar elementos dentro de um vetor. Nos exemplos (5) e (10), havíamos criado os vetores **x** e **y**, cada qual com apenas um elemento. Com a função `c()`, podemos criar um vetor com um número maior de elementos, por exemplo, os nomes das variáveis sociais de nosso *corpus*:

(12)

```
>variaveis.sociais<-c("sexo.genero","faixa.etaria","escolaridade")¶
>variaveis.sociais¶
[1] "sexo.genero" "faixa.etaria" "escolaridade"
```

Em (12), `variaveis.sociais` é o nome do vetor/objeto que contém os argumentos "sexo.genero", "faixa.etaria" e "escolaridade". Se pedirmos ao R que mostre o que é `variaveis.sociais` (linha 2 do exemplo), ele retornará os valores do vetor (linha 3 do exemplo). É importante notar que o nome do vetor poderia ser qualquer outro (por exemplo, "var.sociais", "aa", "x" etc.), e que o número de elementos a serem combinados (três, no exemplo acima) poderia ser dois, quatro, vinte, dez mil...

`getwd()/ setwd()`: `getwd()` é uma função que mostra qual é o diretório de trabalho (wd = *working directory*) atual do R. Quando se deseja buscar um arquivo ou salvar resultados dentro do computador, essa é a pasta que o R usará como referência. Para especificar um diretório de trabalho diferente, usa-se a função `setwd()`, cujo argumento é o caminho completo da pasta dentro do sistema, definido entre aspas. Por exemplo:
(13)

```
setwd("C:/Documentos/dmsocio/dmsocio-SP2010")¶
```

No RStudio, em vez de digitar o caminho completo, pode-se definir o diretório de trabalho na aba *Files*, em uma das janelas da interface. O programa indica o diretório de trabalho atual no topo da janela (Figura 2a). Uma nova pasta pode ser escolhida clicando sobre as reticências no canto direito; na janela que se abrir, indique a localização da nova pasta (Figura 2b). O caminho completo para a nova pasta aparece no topo da aba. Em seguida, clique em *More* e em *Set As Working Directory* (Figura 2c).

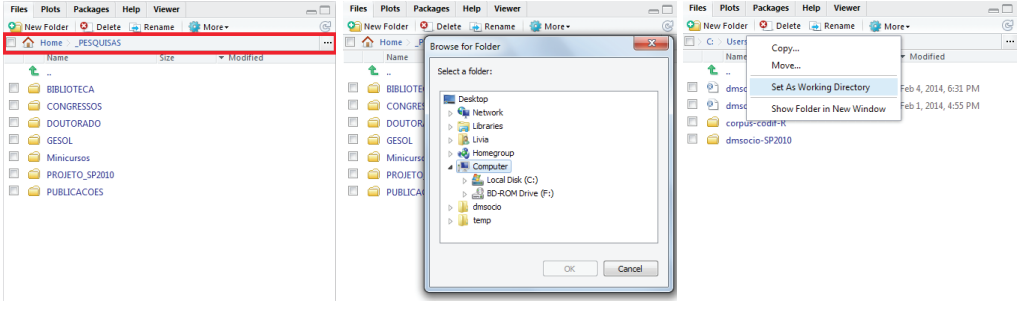


Figura 2 – Aba *Files* do RStudio: (a) diretório atual; (b) navegação para uma nova pasta; (c) definição de nova pasta de trabalho.

`read.table()`: Função para carregar um arquivo de dados no formato de tabela (por exemplo, uma planilha do Excel armazenada em formato .txt) em um vetor do R. Essa função possui diversos argumentos, alguns dos quais com um valor padrão (*default*). Em geral, a existência de um valor padrão ocorre quando há um número limitado de opções para o preenchimento de um argumento – por exemplo, **T** (verdadeiro, do inglês *true*) ou **F** (falso, do inglês *false*). Note que, no caso das funções `c()`e `getwd()/setwd()`, não há um valor padrão para os respectivos argumentos, dado que o número de possibilidades de preenchimento é infinito: para `c()`, pode-se definir quaisquer elementos a serem combinados e em qualquer número; para `getwd()/setwd()`, o caminho do diretório de trabalho depende das pastas existentes no sistema do usuário, que podem ter qualquer nome. Quando uma função possui valor padrão para certos argumentos, pode-se simplesmente não especificá-los, caso em que o R automaticamente assume que o valor do argumento é o padrão. Os argumentos de `read.table()` mais relevantes para nossos propósitos, com seus respectivos valores padrão, são os seguintes:

(14)
```
read.table(file=...,header=F,sep="",quote="",comment.char="#")
```

- O argumento *file*, obrigatório, define o arquivo a ser lido pelo R. Pode-se especificar o caminho completo até ele (por exemplo, C:/Documentos/dmsocio/CodifR-12ent.txt) ou, quando o arquivo está no diretório de trabalho definido com `setwd()`, basta especificar o nome do arquivo (por exemplo, CodifR-12ent.txt). Alternativamente, o argumento de *file* pode ser especificado como `choose.files()` (no Linux, como `file.choose()`); nesse caso, o R abre uma janela de navegação na qual o usuário pode localizar o arquivo desejado e clicar sobre ele;
- *header* especifica se a primeira linha da tabela de dados a ser lida pelo R contém o nome das colunas ou não. O valor padrão é **F** (= falso), mas se a tabela que você deseja carregar no R possui o nome das colunas, esse argumento deve ser definido como **T** (= verdadeiro);
- *sep* define o caractere que separa os campos da tabela (espaços, tabulações – tabs, vírgulas, ponto-e-vírgulas, etc.). Em muitos casos, utilizaremos o tab como separador de campos, representado no R como `"\t"`;
- *quote* define os caracteres usados para citação. Na maior parte dos casos, é melhor defini-lo como `quote=""` (ou seja, nenhum), para evitar problemas na leitura do arquivo pelo R;

- *comment.char* define o caractere usado para comentários. Em linguagens de programação, o caractere de comentário é usado para instruir o programa a ignorar tudo que vem depois dele, de modo que o usuário possa fazer anotações sobre suas linhas de comando. No R, o caractere de comentário padrão é # mas, para nossos propósitos (e a depender das normas de transcrição de seu *corpus*), provavelmente você preferirá defini-lo como `comment.char=""`.

Os materiais de treino para uso do pacote `dmsocio` incluem uma tabela de dados chamada "CodifR-12ent.txt". Ela pode ser carregada no R através da linha de comando exemplificada em (15), que instrui o programa: (i) a abrir a janela de navegação (`file=choose.files()`); (ii) que a primeira linha da tabela é o nome das colunas (`header=T`); (iii) que o caractere que separa os campos é o tab (`sep="\t"`); (iv) que não há caractere de citação nem de comentário (`quote=""; comment.char=""`); e (v) que tais dados devem ser armazenados no vetor chamado `dadosR` (`dadosR<-`).

(15)

```
dadosR<-read.table(file=choose.files(), header=T, sep="\t",
quote="", comment.char="")¶
```

Além dessas funções, é útil descrever outras cinco adicionais: `scan()`, `grep()`, `gsub()`, `cat()` e `write.table()`. Embora você não precise conhecê-las para utilizar o pacote `dmsocio`, trata-se de funções que são empregadas dentro das rotinas dos *scripts* que compõem o pacote e que podem ser úteis em suas próprias explorações de arquivos de dados.

`scan()`: Função para carregar dados de um arquivo de texto (por exemplo, um arquivo de transcrição) em um vetor. Assim como a função `read.table()`, `scan()` possui diversos argumentos. Os mais relevantes para nossos propósitos são discriminados em (16) e exemplificados em (17):

(16)

```
scan(file=..., what="double(0)", sep="", comment.char="")
```

(17)

```
> FabianaB<-scan(file=choose.files(),what="char",sep="\n")¶
> FabianaB ¶
```

- *file* é obrigatório e especifica o caminho do arquivo a ser carregado, de modo semelhante à função `read.table()`. A especificação `file=choose.files()` abre a janela de navegação para que o arquivo seja localizado manualmente;

- *what* especifica o tipo de dados que o arquivo contém: números (`what="double(0)"`) ou caracteres de texto (`what="char"`) – este último, mais frequentemente, será o nosso caso;
- *sep* define o caractere que separa os elementos do vetor. O *default* `sep=""` divide o texto em palavras e guarda cada uma como um elemento do vetor; `sep="\n"` divide o texto por parágrafos (cada parágrafo se torna um elemento do vetor);
- *comment.char* define o caractere usado para comentários, de modo semelhante à função `read.table()`.

No exemplo em (17), pedimos ao R que abra a janela de localização; podemos então clicar sobre o arquivo "SP2012-009-F26SEL-FabianaB.txt", na pasta "dmsocio-2010", para carregar essa transcrição no vetor `FabianaB`. Ao digitar o nome do objeto recém-criado no console, o R retorna seus elementos (a transcrição).[3]

É importante apontar que qualquer operação realizada com o vetor (por exemplo, as funções `grep()` e `gsub()` discutidas abaixo) **não** altera o arquivo original, uma vez que os dados estão armazenados na sessão corrente do R. Se desejamos salvar as alterações ou resultados, podemos usar as funções `cat()` ou `write.table()` (ver adiante).

`grep()`: Função que realiza a busca de padrões dentro de um vetor.

(18)

```
grep(pattern, x, ignore.case=F, value=F)
```

(19)

```
> chopes<-grep("chopes", FabianaB, ignore.case=T, value=F)¶
> chopes ¶
```

- *pattern* se refere ao padrão a ser buscado (a subseção 2.3 tratará disso mais detalhadamente);
- *x* define o vetor em que a busca deve ser realizada;

3 No exemplo 17, se os caracteres acentuados como "ã" e "ó" não aparecerem corretamente no console após digitar `"FabianaB"`, tente uma das soluções seguintes: (i) clique sobre *Tools* no menu superior e, em seguida, em *Global options*; na janela que aparecer, no menu à esquerda, clique em *General* e mude o *Default text encoding* para *Windows-1252*; ou (sobretudo para usuários de Linux) (ii) utilize os arquivos em formato .txt – UTF-8 da pasta "dmsocio", mantendo o *Default text encoding* como UTF-8.

- *ignore.case* define se a diferença entre letras maiúsculas e minúsculas (por exemplo, entre "SP" e "sp") deve ser ignorada ou não. O *default* é **F**, ou seja, o R não deve ignorar a diferença maiúsculas e minúsculas, mas o usuário pode definir o argumento como **T** caso a diferença seja irrelevante;
- *value* retorna a localização das ocorrências quando **F**, ou as próprias ocorrências quando **T**.

No exemplo em (19), a função `grep()` instrui o programa a buscar ocorrências do item lexical "chopes" no vetor `FabianaB`, e guardar os resultados no vetor `chopes`. Como `value=F`, o resultado é a sua localização dentro do vetor.

`gsub()`: Função que localiza padrões em um vetor e os substitui por outro valor. Os argumentos *pattern* e *x* são definidos de modo semelhante à função `grep()`. O argumento adicional *replacement* define o valor que deve substituir o padrão.

(20)

```
gsub(pattern, replacement, x, ignore.case=F)
```

(21)

```
> FabianaB.falantes<-gsub("D1","Doc",FabianaB)¶
> FabianaB.falantes<-gsub("S1","Inf",FabianaB.falantes)¶
> FabianaB.falantes¶
```

No exemplo (21), substituimos todas as ocorrências de "D1" por "Doc", e de "S1" por "Inf". O resultado pode ser visualizado com a chamada do vetor `FabianaB.falantes`.

`cat()/write.table()`: Funções que permitem salvar, respectivamente, vetores e *dataframes* em arquivos. Em ambas, *x* se refere ao objeto a ser salvo, *file* ao nome do arquivo (junto com sua extensão, por exemplo, .txt), e *sep* define o tipo de separador dos elementos (espaços `" "`, parágrafo `"\n"`, tab `"\t"`, etc.). `write.table()` possui o argumento adicional *append*, que define se novos valores devem ser anexados ao fim da tabela (**T**) ou não (**F**).

(22)

```
cat(x, file="",sep="")
```

(23)

```
> cat(FabianaB.falantes, file = "FabianaB-part.txt", sep="\n")¶
```

(24)
```
write.table(x, file="", append=F, sep="")
```

No exemplo em (23), a função `cat()` salva o vetor recém-criado `FabianaB.falantes` no arquivo "FabianaB-part.txt" no atual diretório padrão.

2.3. Expressões regulares

Expressões regulares são sequências de caracteres que especificam padrões de busca. Através do uso de caracteres especiais, pode-se definir um padrão que contém não apenas grafemas específicos como "a", "bol" ou "chopes", mas também a localização desses caracteres dentro de uma palavra ou expressão, caracteres opcionais ou conjuntos de caracteres que podem ocupar uma determinada posição. O Quadro 1 apresenta os caracteres especiais que serão mais utilizados em nossas definições de padrões a serem buscados nas entrevistas sociolinguísticas.

[]	Trata o que está dentro como uma classe de caracteres
|	"Ou"
.	"Qualquer caractere"
*	"Zero ou mais ocorrências da expressão regular precedente"
+	"Uma ou mais ocorrências da expressão regular precedente"
?	"Zero ou uma ocorrência da expressão regular precedente"
\\b	Fronteira de palavra
\\d	Qualquer dígito
\\s	Qualquer espaço
\\	Caracteres de escape

Quadro 1 – Caracteres especiais para definição de expressões regulares.

Tomemos como exemplo dois pequenos trechos extraídos da entrevista com o informante MauricioB, parte do *corpus* do Projeto SP2010 e de nossa amostra para treino do uso das funções do pacote dmsocio, em que o informante fala sobre os problemas de São Paulo. Em (25), "D1" representa a fala do documentador e "S1" a fala da informante.

(25)

```
 1  S1 já já ouvi/ já vi várias pesquisas tal porque o que que
 2     está apavorando hoje o morador de São Paulo...
 3     é a maldita violência né?
 4  D1 uhum
 5  S1 é a maldita violência então isso a pessoa... pensa
 6     duas vezes três vezes por mais que apaixonada seja
 7  S1 ela vai falar não eu quero um lugar pra... pra ter
 8     mais tranquilidade
 9  S1 mas em contrapartida você também lê toda hora
10     que... os lugares mais calmos do interior já estão
11     sendo
12  S1 vítimas aí de arrastão e de assalto de tudo quanto
13     que é tipo
14  D1 aham
15  S1 então não sei até que... até que ponto valeria a pena
```

[...]

```
16  S1 poluição está
17  S1 São Paulo hoje em dia... fica três dia sem chover
18     você começa a lacrimejar já começa
19  S1 a travar tua garganta
20  S1 (quer dizer)
21  D1 você sofre com isso assim com essas coisas?
22  S1 ah eu não sentia até tempo atrás hoje em dia eu sinto
```

O mesmo trecho pode ser carregado em um vetor no R com a linha de comando em (26). Ao abrir a janela de navegação, escolha o arquivo "Ex26-MauricioB.txt".

(26)

```
> MauricioB<-scan(file=choose.files(),what="char",sep="")¶
> MauricioB¶
```

Imagine que um pesquisador esteja interessado em buscar todas as ocorrências da realização de /e/ nasal, em posição pré-consonantal, que não ocorram no início, nem no final de palavra. Se o pesquisador tiver acesso a uma transcrição fonética do *corpus*, sua tarefa será bem mais simples. No entanto, na maior parte dos casos, o sociolinguista lida com transcrições ortográficas das gravações. A

tarefa aqui é definir possibilidades ortográficas (do mundo da escrita) que representam um segmento fônico (do mundo da fala).

No trecho acima, as palavras-alvo são as ocorrências da palavra "violência" (linhas 3 e 5), "pensa" (linha 5), "sendo" (linha 11), "sentia" e "tempo" (linha 22). Uma busca apenas pela sequência "en" encontraria as palavras "pensa", "sendo" e "sentia", mas deixaria de encontrar as ocorrências da palavra "violência", que contém o acento circunflexo, e "tempo", em que /e/ nasal é representado graficamente por "em". Ao mesmo tempo, a busca por "en" também encontraria as palavras "então" e "pena", que não fazem parte do conjunto: a primeira pelo fato de /e/ nasal estar em posição inicial da palavra e a segunda por ser seguido de vogal. Por outro lado, a busca pela sequência "em" teria problemas semelhantes: apesar de encontrar, corretamente, a ocorrência da palavra "tempo", a busca também retornaria com as palavras "em" e "sem", em que /e/ nasal se encontra no início e no final da palavra, respectivamente.

Desse modo, queremos que a busca inclua: (i) não apenas o grafema "e", mas também suas versões acentuadas ("ê" e "é"); (ii) não apenas os grafemas "e", "ê" e "é" seguidos de "n", mas também aqueles seguidos de "m"; (iii) que as sequências dos grafemas "en", "ên", "én", "em", "êm", "ém" sejam seguidas de consoantes ("b", "c", "ç", "d" etc.) e (iv) que tais sequências não ocorram nem no início nem no final da palavra.

Caso resolvêssemos fazer uma tal busca em um editor de texto (como o Word ou Bloco de Notas), seriam necessárias múltiplas definições da sequência de caracteres desejada – contando as seis combinações de "e"/"ê"/"é" com "n"/"m" acima, com todas as possíveis combinações de consoantes seguintes, o número total de definições de sequências certamente estaria na casa de centenas. Com o emprego de expressões regulares e dos caracteres especiais do Quadro 1, é possível definir o padrão desejado de modo bastante econômico. Vejamos:

Os caracteres especiais `[]` tratam aquilo que está dentro como uma classe de caracteres. Para definir que desejamos ocorrências de "e", "ê" ou "é", podemos criar uma classe de caracteres `[eêé]`[4]. De modo semelhante, podemos definir a presença de "n" ou "m" como `[nm]` e a presença de uma consoante como `[bcç-dfgjklpqrstvxz]`. Veja que essa última classe de caracteres não inclui "h" (pois faria com que palavras como "nenhum" entrassem no conjunto de palavras-alvo), tampouco "n" ou "m" (pois buscaria sequências de "nn" e "mm", inexistentes no português). Até o momento temos, portanto, a seguinte sequência de caracteres:

(27)

```
[eêé][nm][bcçdfgjklpqrstvxz]
```

4 A mesma sequência de caracteres poderia ser notada como [elêlé], utilizando o caractere especial | (ou). No entanto, como [] cria uma classe de caracteres, o uso de |, nesse caso, seria redundante.

Note que a definição de que /e/ nasal não pode ocorrer em final de palavra já está incluída na sequência em (27), uma vez que a presença de uma consoante após [nm] garante que o segmento não estará no final na palavra.

Agora precisamos definir que a sequência em (27) não deve ocorrer em início de vocábulo. Para tanto, podemos usar os caracteres especiais `\\b`, que delimitam fronteiras de palavra. Se a sequência não pode ocorrer no início, isso significa que há algum caractere – qualquer que seja – após o seu início, o que é representado pelo ponto final (.) (qualquer caracter); além disso, pode haver apenas um caractere antes da sequência (como na palavra "t-emp-o"), ou pode haver mais (como na palavra "viol-ênc-ia"), o que se representa pelo caractere especial + (uma ou mais ocorrências da expressão regular precedente, nesse caso, o ponto final). Podemos, portanto, atualizar nossa definição como em (28), e inseri-la na função `grep()` para localizar tais ocorrências no vetor `MauricioB` (29):

(28)

```
\\b.+[eêé][nm][bcçdfgjklpqrstvxz]
```

(29)

```
> grep("\\b.+[eêé][nm][bcçdfgjklpqrstvxz]", MauricioB, value=T)¶
```

A expressão regular em (28-29) identifica corretamente as ocorrências de /e/ nasal "violência", "pensa", "sendo", "sentia" e "tempo", ao mesmo tempo em que não identifica as palavras "então", "pena", "também", "em" e "sem" como parte do conjunto de palavras-alvo. Ela cumpre os dois critérios que devem ser atendidos quando se define uma expressão regular: (i) que abarque **todos** os casos do padrão buscado; e (ii) que abarque **somente** os casos do padrão buscado.

Alguns exemplos adicionais (incluindo a utilização de caracteres especiais não discutidos nessa subseção) serão apresentados adiante.

3. PREPARAÇÃO

3.1. Regra #1: Conheça sua variável!

O exemplo fornecido na subseção anterior pode parecer idiossincrático: por que buscar "ocorrências da realização de /e/ nasal, em posição pré-consonantal, que não ocorram no início e nem no final de palavra?" Tal definição faz parte do contexto variável para o estudo da realização de /e/ nasal no português paulistano como um monotongo (por exemplo, [fa.ˈzẽ.da]) ou ditongo (por exemplo, [fa.ˈzẽ̃j.da]) (OUSHIRO, 2013). Na análise qualitativa do *corpus*, notou-se que

as ocorrências de /e/ nasal em sílabas átonas e em posição inicial (por exemplo, "então", "em" e "engravidar") ou final (por exemplo, "fazem" e "vagem") eram quase que invariavelmente realizadas como [ĩ] – e, no caso das pós-tônicas finais, por vezes desnasalizadas [i]. Essas ocorrências podem ser entendidas como manifestação de outras variáveis, a saber, o alçamento de vogais pré e pós- tônicas (/e/ → [i]) (Cf. por exemplo, VIEGAS, 1987; BATTISTI, 1993; CELIA, 2004; TENANI; SILVEIRA, 2008) e a desnasalização (/ẽ/ → [i]) (Cf. por exemplo, VOTRE, 1978; GUY, 1981; BATTISTI, 2002; SCHWINDT; SILVA, 2009). Em monossílabos tônicos (por exemplo, "tem" e "vem") ou oxítonas (por exemplo, "também" e "armazém"), o segmento é invariavelmente realizado como ditongo [ẽj̃]. Em posição pré-vocálica (por exemplo, "pena") ou quando há assimilação de "nd" (como em "fazendo"), o segmento é invariamente realizado como monotongo.

Todas essas observações advêm da leitura da bibliografia relevante e de uma análise qualitativa cuidadosa do *corpus*, com vistas à definição do *contexto variável* (LABOV, 1969): qual é o conjunto das variantes e em quais contextos elas podem ocorrer (mesmo que não tenham ocorrido)?

Nesse ponto, vale destacar o argumento lançado inicialmente: para tarefas repetitivas, mecânicas e previsíveis, o R é uma ferramenta poderosa na redução do tempo empregado para o seu cumprimento, mas o programa por si só não pode fornecer respostas a perguntas que cabem ao linguista e que dependem de seu conhecimento especializado. Em outras palavras: o programa não pensa por você. Ele pode realizar tão somente as tarefas que você definir.

Na aplicação das funções `identificacao()` e `extracao()`adiante, discutiremos exemplos com outras três variáveis do português: (i) o emprego de diminutivos; (ii) a pronúncia de (-r) em coda silábica e (iii) o uso dos pronomes de primeira pessoa do plural "nós" e "a gente". Por ora, deixemos de lado a definição das expressões regulares correspondentes; antes disso, é necessário refletir sobre as questões que justificam o interesse por esses fenômenos linguísticos e tomar decisões quanto ao contexto variável.

Em um levantamento preliminar sobre a fala *gay* no português (MENDES, 2011), alguns dos informantes disseram associar o uso de diminutivos (por exemplo, "casinha" e "amiguinho") com a fala das mulheres e que, quando empregado de modo "exagerado", um falante do sexo masculino poderia ser percebido como *gay*. Desse modo, Mendes (2012) investigou se de fato há diferenças no emprego de diminutivos de acordo com o gênero dos falantes – homens e mulheres, *gays* e heterossexuais. Tal empreitada, no entanto, conduziu a um desafio metodológico: como definir o contexto variável para o uso de diminutivos? Em português, os diminutivos podem ocorrer em qualquer substantivo ("casa", "casinha"), adjetivo ("perto", "pertinho"), advérbio ("falar rápido", "falar rapidinho") e gerúndio ("tá chovendo", "tá chovendinho"). Extrair *todas*

as ocorrências dessas classes de palavras e codificá-las de acordo com a presença ou não de sufixo de diminutivo parecia ineficaz. Mendes então decidiu coletar apenas as ocorrências de diminutivos no *corpus* e, sabendo o número total de palavras por entrevista sociolinguística, calculou o número de ocorrências por mil palavras. Esse índice foi usado na comparação da frequência de diminutivos na fala de homens e mulheres, *gays* e heterossexuais.

A pronúncia de /r/ em português, tanto em posição de ataque quanto em posição de coda silábica, possui múltiplas variantes: vibrante múltipla, tepe, aproximante retroflexa, fricativa velar e glotal (surda e sonora), apagamento. As diferentes realizações também costumam ser associadas às diferentes variedades regionais e sociais. Em posição de coda silábica, as variantes mais comumente associadas ao português paulistano são o tepe e o retroflexo. Na capital paulista, o retroflexo tradicionalmente se associa aos falantes "caipiras", mas as suas indexicalidades parecem estar se expandindo para abarcar uma identidade de morador de periferia urbana. Em seu estudo sobre a realização de /-r/ em coda, Oushiro e Mendes (2013) incluíram as ocorrências tanto em contexto medial (por exemplo, "porta") quanto final (por exemplo, "mulher") e excluíram ocorrências de (-r) em fronteira de palavra seguido por vogais (por exemplo, "por isso"), uma vez que em tais ocorrências o segmento /r/ passa a ocupar a posição de ataque silábico ([po.ˈri.sʊ]).

Por fim, um dos principais interesses em analisar o emprego variável do pronome de primeira pessoa do plural é o processo de mudança linguística com o surgimento de um novo pronome, "a gente", que se encontra em alternância com o pronome mais antigo "nós" (ZILLES, 2005). Para esse caso, cabe ao pesquisador decidir se analisará somente os casos nominativos, na posição de sujeito ("nós fomos/foi", "a gente fomos/foi"), ou se também analisará os pronomes em outras posições sintáticas, no caso acusativo ("ele nos viu", "ele viu a gente"), como possessivo ("o nosso bairro", "o bairro da gente"), etc. Tais decisões devem ser tomadas em consonância com as questões que norteiam a pesquisa, em diálogo com a literatura relevante e de acordo com os dados disponíveis ao pesquisador.

3.2 Regra #2: Conheça seu corpus

A definição de expressões regulares para busca de ocorrências no *corpus* depende fundamentalmente das normas de transcrição empregadas, que geralmente diferem entre grupos de estudos e pesquisadores (cf. CASTILHO; PRETTI, 1986 para o Projeto NURC, TENANI; GONÇALVES, s/d para o Projeto ALIP, MELLO; RASO, 2009 para o C-ORAL-BRASIL, MENDES; OUSHIRO, 2013 para o Projeto SP2010).

Um princípio geral que norteia a estipulação de normas de transcrição é a representação da fala por meio escrito, de forma fiel à língua oral, mas inteligível, de modo que o material coletado possa ser mais facilmente analisado. Entretanto, as normas de transcrição também levam em conta os interesses específicos dos pesquisadores e seus respectivos grupos de pesquisa, tendo em vista as análises que serão desenvolvidas a partir do material coletado (análises fonéticas, morfológicas, sintáticas, textuais, etc.)

Diferentes normas de marcação de pausas, alongamento de vogais, apagamento e assimilação de segmentos, hesitações, turnos dos falantes, dados contextuais, presença de cabeçalho, etc. fazem com que a busca de uma mesma expressão regular retorne resultados diferentes, a depender das convenções adotadas. Tomemos as sentenças em (30), todas inventadas, para examinar essa questão.

(30)
a. na 3ª, 4ª série, eu era muito bagunceira
b. daí ele gritou assim... "abre a PO::::rta!"
c. eles vão e [barulho de porta batendo] fazem
d. eu num vou falá(r) p(r)a ele que ele (es)tá errado né?

Um pesquisador interessado em analisar a pronúncia de (-r) em coda silábica deve coletar, desses pequenos exemplos, os itens "3ª" e "4ª" em (30a), "PO::::rta" em (30b), e "falá(r)" em (30d), e ignorar a ocorrência de "porta" em (30c), que se refere à anotação de um dado contextual e, portanto, não se trata de um dado linguístico. Uma expressão regular que define a variável (-r) em coda como [aáâeéêiíoóôuú]r[bcçdfgjklmnpqstvxwz] – ou seja, a sequência de uma vogal, o grafema "r" e uma consoante – não seria capaz de identificar as ocorrências corretamente – de fato, ela só identificaria a ocorrência em (30c), que é justamente aquela que não interessa na análise.

A depender das normas de transcrição do *corpus* com que você está trabalhando, pode ser mais interessante fazer certos ajustes *antes* de realizar a busca automática por ocorrências, ao invés de tentar dar conta de todas as possibilidades de transcrição ortográfica. Para fazer isso, vamos primeiro combinar os dados de (30) em um vetor chamado `exemplo.R`:

(31)

```
 > exemplo.R<-c("na 3ª, 4ª série, eu era muito bagunceira",
"daí ele gritou assim... "abre a PO::::rta!"",
"eles vão e [barulho de porta batendo] fazem",
"eu num vou falá(r) p(r)a ele que ele (es)tá errado né?")¶
```

Nesse exemplo, sabemos de antemão que as palavras "3ª" e "4ª" contêm ocorrências de (-r) em coda silábica; no entanto, em um *corpus* maior, o pesquisador pode primeiro querer inspecionar quais e quantas ocorrências contêm dígitos (`\\d`) não transcritos por extenso. Isso pode ser feito através do uso de uma expressão regular com o `grep()`:

(32)

```
> grep("\\d", exemplo.R, value=T) ¶
[1] "na 3ª, 4ª série, eu era muito bagunceira"
```

O pesquisador agora sabe que existem as palavras "3ª" e "4ª" que devem ser substituídas por sua transcrição por extenso. Depois disso, podemos instruir o R a substituir os sinais `:()[]` – e qualquer conteúdo dentro de `[]` – por `""` (ou seja, nada, o que equivale a apagá-los). Isso pode ser feito em uma única linha de comando, separando cada valor a ser apagado com | (ou). Aqui, no entanto, deve-se tomar uma medida adicional com os sinais de parênteses e colchetes, que fazem parte de um conjunto de caracteres com significado especial no R. Esses incluem: ^ $ * . + ? ! { } – alguns descritos na subseção 2.3 (ver também Gries, 2009, p. 81). Para que sejam entendidos literalmente, e não com o seu significado especial, devemos notá-los junto ao símbolo de escape `\\` – por exemplo, `\\(`. A mesma função `gsub()` pode ser utilizada para substituir "3ª" por "terceira" e "4ª" por "quarta". Note que em (33) abaixo, cada alteração é armazenada em um novo vetor (`exemplo.R1`, `exemplo.R2`, `exemplo.R3`) e que cada nova alteração é realizada no vetor mais atualizado.

(33)

```
> exemploR1<-gsub(":|\\(|\\)|\\[.+\\]", "", exemplo.R)¶
> exemplo.R2<-gsub("3ª", "terceira", exemplo.R1)¶
> exemplo.R3<-gsub("4ª", "quarta", exemplo.R2)¶
> exemplo.R3¶
```

Todas essas considerações também podem ser levadas em conta caso você esteja na fase de transcrição do *corpus*: quais normas facilitarão, posteriormente, o tratamento de dados[5]?

Muitos projetos de estudos sociolinguísticos costumam disponibilizar certas informações da ficha social do informante (sexo, idade, escolaridade, etc.) na forma de cabeçalho. A Figura 3 é um exemplo retirado da Amostra Censo/1980

5 As normas de transcrição do Projeto SP2010 foram definidas com essa preocupação em mente: a facilidade de tratamento de dados com o R. O manual de transcrições desse projeto (MENDES; OUSHIRO, 2013) pode ser acessado no endereço <http://projetosp2010.fflch.usp.br/producao-bibliografica>, no link *Manual de Transcrições*.

do Projeto PEUL[6] (PAIVA; SCHERRE, 1999). Além de dados institucionais, o cabeçalho identifica o informante, sua idade, escolaridade, bairro de residência e profissão. Esses dados (ou parte deles) poderão ser extraídos automaticamente com o R, para codificação das variáveis sociais de interesse na análise.

Universidade Federal do Rio de Janeiro - UFRJ
Centro de Letras e Artes – CLA
Faculdade de Letras
Departamento de Lingüística e Filologia
Programa de Estudos sobre o Uso da Língua - PEUL
Banco de Dados do PEUL/UFRJ
AMOSTRA CENSO/1980

Falante: 01 Sam
Idade: 18 anos
Escolaridade: Fundamental 1
Bairro: Santa Cruz
Profissão: ajudante de pedreiro

(gravação feita em ambiente com eco)
E- (fala cortada) (ruídos) (inint.) ("com sua mãe"). Com quem que você se dá melhor, dos seus irmãos?
F- Meus irmão, eu me dou melhor [com]- com o meu irmão abaixo de mim. (est)
E- Ele é muito mais novo?
F- Não, ele tem (falam) uns quinze anos. (est)

Figura 3 – Exemplo de transcrição do PEUL, com cabeçalho.

3.3. Arquivos

Antes de discutir a aplicação das funções do pacote `dmsocio`, alguns cuidados adicionais são necessários na preparação do *corpus*. Ainda que o programa R seja capaz de "ler" diversos tipos de arquivo, as funções que serão discutidas na próxima seção requerem que os arquivos de transcrição e o de dados estejam salvos no formato .txt (*Texto sem formatação* no Word e *Texto delimitado por tabulador* no Excel). Esse é, de fato, o formato mais flexível para o tratamento de dados textuais.

Em muitos casos, o sociolinguista dispõe de transcrições em formato .doc, .docx, .odt ou .pdf. No caso dos primeiros, basta abrir os arquivos em um editor de texto como o Word ou o Writer e salvá-los no formato .txt. Para arquivos em .pdf, há uma série de programas gratuitos na Internet que realizam a conversão automática de arquivos .pdf para .txt.

6 Disponível em <http://www.letras.ufrj.br/peul/cen80texto.html>. Acesso em: 30 jan. 2014.

Como se mencionou na subseção anterior, pode ser útil incluir um cabeçalho com informações sociais de cada informante, caso o seu *corpus* não tenha um. Os tipos de informação e a ordem em que aparecem podem variar, mas é importante manter consistência entre os arquivos: se a idade do informante aparece na quinta linha, ela deve aparecer na quinta linha para todos os informantes e transcrições. O mesmo vale para outras informações: se a fala do documentador é marcada por "E- " ("e" maiúsculo, traço, espaço – ver Figura 3), isso deve ocorrer ao longo de todas as transcrições. A regra mais básica aqui é: não importa o que se faça, faça consistentemente.

Para os arquivos de transcrição (mas não para os arquivos de dados), recomenda-se apagar as marcas de tabulação, se houver. No R, isso pode ser feito com a função `gsub()` (ver exemplo 33 acima), utilizando "\t" para representar os tabs. Isso é necessário pois a função `extracao()` gera uma tabela de dados separados por tabs; caso o arquivo original já contenha essas marcas, o de dados final pode acabar contendo algumas colunas adicionais indesejadas.

4. FUNÇÕES `IDENTIFICACAO()`, `EXTRACAO()`, `AMOSTRAGEM()`

Esta seção discute a aplicação de três funções do pacote `dmsocio`, desenvolvidas especificamente para o tratamento de dados sociolinguísticos. Os exemplos discutidos utilizam as transcrições do *minicorpus* "dmsocio-SP2010" (12 entrevistas sociolinguísticas do Projeto SP2010) e os arquivos codificados na pasta "dmsocio-codif-R". Para que as funções estejam disponíveis em cada sessão do R, é necessário rodar uma linha de comando que as carrega na memória:

(34)

```
> source("~/dmsocio.R")¶
```

em que "~" corresponde ao caminho completo da pasta em que se encontra o arquivo "`dmsocio.R`" no seu computador. Alternativamente, se você estiver conectado à internet pode rodar a seguinte linha de comando:

(35)

```
>source("http://tinyurl.com/dmsocio")¶
```

Essa segunda opção lê o script dmsocio.R diretamente do portal do Projeto SP2010. A vantagem dessa segunda opção é ter acesso a versões mais recentes, com todas as atualizações.

4.1. Função `identificacao()`

A função `identificacao()` busca as ocorrências de um determinado padrão em um *corpus* e gera novos arquivos de transcrição com marcações de ocorrências e novo nome, na mesma pasta em que estão os arquivos originais. Ela possui diversos argumentos, que são descritos a seguir:

`padrao`	Obrigatório. Definição da sequência de caracteres (expressão regular) que identifica as variantes da variável. Deve ser especificado entre aspas;
`simbolo. marcacao`	Opcional. Definição dos símbolos que serão usados para identificar as ocorrências da variável no *corpus*. *Default* = "<>";
`posicao.marcacao`	Obrigatório. Lógico (T ou F). Se T, a marcação é colocada após o padrão. Se F, a marcação é colocada antes do padrão. Default=T. Nota: Para posicao.marcacao=F, a definição da variável deve começar com \\b (levando em conta o início da palavra);
`ignorar.linhas`	Opcional. Vetor com identificação dos participantes, cujas falas devem ser ignoradas na busca do padrão. *Default*=NULL.
`stoplist`	Opcional. Vetor com palavras que devem ser ignoradas na busca do padrão. *Default*=NULL.
`novos.arquivos`	Opcional. Sequência de caracteres a ser adicionada ao nome dos arquivos originais, para diferenciar os arquivos com marcações de ocorrência na mesma pasta. Deve ser especificado entre aspas. *Default*="marcacoes".

Note, primeiramente, que quatro dos argumentos da função `identificacao()` são opcionais: `simbolo.marcacao`, `ignorar.linhas`, `stoplist` e `novos.arquivos`. Se tais argumentos não são especificados, a função utiliza cada valor *default* respectivo.

Essa função requer que se especifique a pasta em que se encontram as transcrições originais, nas quais o padrão será buscado. Como visto em 2.2, isso pode ser feito com a função `setwd()` ou através da aba *Files* do RStudio.

A seguir, discutimos exemplos da aplicação da função na localização de três variáveis: diminutivos, (-r) em coda silábica e pronome de primeira pessoa do plural.

Exemplo 1: diminutivos

O caso mais simples de uso da função `identificacao()` é aquele em que se especifica somente o padrão a ser buscado. Suponha que queremos investigar o uso dos diminutivos no português, em seu caso mais comum: o diminutivo sintético com sufixo "-inh-"[7]. Precisamos, primeiro, definir para o R o que é um diminutivo através de uma expressão regular. Antes de continuar a leitura, imagine como você definiria tal expressão, usando os caracteres especiais vistos em 2.3.

A busca deve incluir palavras terminadas em "-inho", "-inha", "-inhos" ou "-inhas". Como "o" e "a" se alternam, podemos especificar `inh[oa]`; como o morfema –s de plural é opcional, podemos usar o caractere especial ? (zero ou uma ocorrência da expressão regular precedente): `inh[oa]s?`; por fim, queremos que essa sequência esteja no final da palavra: `inh[oa]s?\\b` é o padrão a ser buscado. Inserindo-o na função `identificacao()` (apenas com os argumentos obrigatórios), temos:

(36)

```
>identificacao(padrao="inh[oa]s?\\b", posicao.marcacao=T)¶
```

O tempo para rodar cada função pode variar consideravelmente (de poucos segundos a alguns minutos), a depender principalmente do tamanho do *corpus* e do número de ocorrências do padrão buscado. Enquanto a função está sendo rodada, aparece um símbolo vermelho (STOP) no canto superior direito do console no RStudio. A função terá terminado quando esse símbolo desaparecer e aparecerem os sinais >| no console.

Os novos arquivos de transcrição, com a adição de "marcacoes-" ao nome original, podem ser visualizados na mesma pasta em que estavam os arquivos originais ou na aba *Files* do RStudio. Examinemos os resultados no arquivo marcacoes-SP2012-009-F26SEL-FabianaB.txt no próprio RStudio: clique sobre o nome do arquivo, que abrirá na janela *Source* (dos *scripts*). Em seguida, digite [CTRL]/[Command]+F para abrir a janela de *Localizar/Substituir* (respectivamente, *Find* – o campo com a lupa – e *Replace*), e digite <> no campo *Localizar* (Figura 4). Clique sobre *Next* para visualizar as próximas marcações.

7 Deve-se lembrar que diminutivos também têm a forma analítica (por exemplo, homem pequeno). Para os sufixos sintéticos, Cunha & Cintra (2007) listam 22 sufixos: -inho(a); -zinho(a); -ino(a); -im; -elho(a); -ejo; -ilho(a); -acho(a); -icho(a); -ucho(a); -ebre; -eco(a); -ico(a); -ela; -ete; -eto(a); -ito(a); -zito(a); -ote(a); -isco(a); -usco(a); e -ola. De todos eles, no entanto, o sufixo *–inh-* é o mais produtivo.

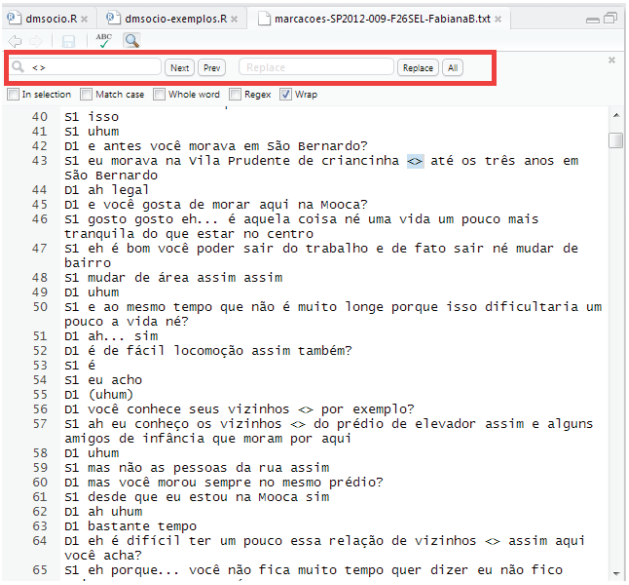


Figura 4 – *Localizar/Substituir* no RStudio

Nesse arquivo, encontramos a marcação de 75 ocorrências: "criancinha" na linha 43, "vizinhos" nas linhas 56, 57, 64, "minha" e "tinha" na linha 76, etc. O R localizou, corretamente, todas as ocorrências de palavras que terminam com "-inho, -inha, -inhos, -inhas". No entanto, deparamo-nos com dois percalços: o primeiro é que o R localizou ocorrências na fala de D1, que é o documentador (por exemplo, nas linhas 56 e 64). Em geral, em estudos sociolinguísticos, estamos interessados em analisar a falar do informante e não nossas próprias falas. Segundo, o R localizou ocorrências de palavras terminadas em -inh- que não são diminutivos, como "vizinhos", "minha" e "tinha".

Vamos lidar com o primeiro deles: o argumento `ignorar.linhas` da função `identificacao()` serve justamente para especificar ao R quais linhas da transcrição devem ser ignoradas na identificação das ocorrências. No *corpus* do Projeto SP2010, o documentador é sempre identificado por "D1", e possíveis outros falantes que participam da gravação são identificados por "S2", "S3" etc. (lembre-se da regra #2: conheça seu *corpus*!). Nas transcrições do *minicorpus*, há entrevistas com até cinco falantes (portanto, até "S5"), além de D1 e S1. Além disso, há linhas separadas para Dados Contextuais e tópicos do Roteiro, que tampouco contêm falas do informante, e as linhas do cabeçalho. O argumento `ignorar.linhas` requer um vetor que especifica quais linhas devem ser ignoradas, o que pode ser realizado com a função `c()`. Em outros *corpora*, é claro, esse vetor seria definido de modo diferente.

(37)

```
>ignorar<-c("D1", "S2", "S3", "S4", "S5", "Dados Contextuais",
"Roteiro","Universidade de", "Faculdade de", "Departamento
 de", "Grupo de", "Projeto SP2010", "(FAPESP)", "Documentador:",
"Informante:", "Arquivos:", "Perfil:", "Sexo/Gênero:", "Idade:",
"Escolaridade:", "Região:", "Zona:", "Geração:", "Renda
 individual:", "Renda familiar:")¶
```

Nosso segundo percalço foi o R localizar ocorrências que não são diminutivos. Trata-se de uma decisão com base em um critério semântico (portanto, não repetitiva, mecânica ou previsível) que o programa, de fato, não poderia fazer por si só, apenas com a definição da expressão regular. Veja também que não é o caso de criar um novo padrão de busca; se instruíssemos o R, por exemplo, a **não** buscar ocorrências que terminem em "minha" ou "tinha", ele deixaria de encontrar possíveis dados como "caminha" (=cama pequena) ou "cartinha", que são diminutivos. Na verdade, o padrão de busca está correto; queremos apenas que o R ignore a ocorrência de certos itens lexicais específicos. Isso pode ser feito através do argumento `stoplist`, um vetor que contém as palavras a serem ignoradas na busca.

Adiante, com a função `extracao()` (seção 4.2), veremos uma maneira mais simples de verificar quais palavras devem entrar na `stoplist` de um modo que não requer a revisão de todas as transcrições. Por ora, fiquemos com as ocorrências apenas da entrevista com FabianaB, para fins de exemplificar a aplicação do argumento `stoplist`. Ao examinar as marcações nessa transcrição, encontramos as seguintes ocorrências de palavras com -inh- que não são diminutivos: "vizinhos" (em que já podemos imaginar que pode haver ocorrências de vizinho, vizinha, vizinhas); "minha", "tinha" e "carinho". Assim, criamos o vetor `pal.stoplist`:

(38)

```
> pal.stoplist<-c("vizinh[oa]s?", "minha", "tinha", "carinho")¶
```

Podemos agora atualizar a função `identificacao()` com os argumentos `ignorar.linhas` e `stoplist`. Antes disso, no entanto, devemos apagar os arquivos "marcacoes-", que já não nos servem mais.

(39)

```
>identificacao(padrao="inh[oa]s?\\b",
        posicao.marcacao=T,
        ignorar.linhas=ignorar,
        stoplist=pal.stoplist)¶
```

Em (39), adicionamos os argumentos `ignorar.linhas` com o vetor `ignorar` e o argumento `stoplist` com o vetor `pal.stoplist` (os argumentos aparecem em diferentes linhas apenas para facilidade de visualização). Ao examinar o arquivo "marcacoes-SP2012-009-F26SEL-FabianaB.txt", vemos que agora as marcações estão corretas: "criancinha" na linha 43, "barzinhos" na linha 78, "senhorzinhos" na linha 133, etc. Os arquivos estão prontos para a codificação da variável dependente, cujas variantes podem ser indicadas dentro dos símbolos <>, de acordo com os códigos estipulados pelo pesquisador[8].

Cabe ainda comentar os demais argumentos da função, que até agora deixamos com a opção *default* ao não especificá-los. O argumento `novos.arquivos` (*default* `"marcacoes"`) pode ser modificado por qualquer outro nome definido pelo usuário, assim como para `simbolo.marcacao` (*default* `"<>"`); para esse último, no entanto, o ideal é que o símbolo escolhido não seja empregado nas normas de transcrição, para que identifique, de modo inequívoco, a ocorrência da variável. Deve-se lembrar, também, que certos caracteres têm significado especial no R (como parênteses, colchetes, chaves, pontuações, etc.); se deseja utilizá-los como símbolo de marcação da variável, devem-se usar os caracteres de escape – por exemplo, para parênteses: `"\\(\\)"`.

O argumento `posicao.marcacao` pode ser definido como antes (=F) ou depois (=T) do padrão buscado. A sua definição como **F** é preferível em padrões que envolvem o início de palavra (diferentemente do caso do diminutivo).

Exemplo 2: (-r) em coda silábica

A variável (-r) em coda silábica exemplifica a aplicação da função `identificacao()` consecutivas vezes, a fim de marcar todas as ocorrências do *corpus*. Trata-se de uma solução possível quando a definição de uma única expressão regular que abarque todas as palavras-alvo se torna muito complicada.

Nesse caso, como na identificação de diminutivos, adotaremos `simbolo.marcacao="<>"`, `posicao.marcacao=T`, `ignorar.linhas=ignorar`, e `novos.arquivos="marcacoes"`; `stoplist`, por sua vez, será mantida como =NULL.

Vejamos inicialmente, como a variável (-r) em coda silábica pode ser definida através de uma expressão regular. Foneticamente, /r/ em coda ocorre após vogais orais, e antes de consoantes (por exemplo, "porta") ou em final de palavra

8 Mendes (2012) codificou as ocorrências de diminutivos do seguinte modo: (i) usos lexicalizados, como *sozinho*, e *barzinho* (que é um tipo específico de bar); (ii) usos metafóricos, como *fumar um cigarrinho* (em que *cigarrinho* não significa um *cigarro pequeno*); e (iii) usos literais, como *a gente vive ali... numa casa bem pequenininha*.

(por exemplo, "mulher"). Ortograficamente, as vogais orais podem ser definidas como uma classe de caracteres [aáâeéêiíoóôuú], e as consoantes como [bcçdfgjklmnpqstvwxz]. Note que na classe das consoantes não se incluem "h" nem "r". Podemos, portanto, definir ocorrências de /r/ em coda seguidas de consoante como:

(40)

```
[aáâeéêiíoóôuú]r[bcçdfgjklmnpqstvwxz]
```

Se desejamos que a marcação da ocorrência apareça após a palavra (por exemplo, "porta <>"), e não após a sequência (por exemplo, "port <>a"), é necessário especificar o fim da palavra \\b como parte da expressão regular, que pode ocorrer zero, um ou mais caracteres ".*?" após a sequência VrC:

(41)

```
[aáâeéêiíoóôuú]r[bcçdfgjklmnpqstvwxz].*?\\b
```

```
>identificacao(padrao="[aáâeéêiíoóôuú]r[bcçdfgjklmnpqstvwxz].*?\\b",
     simbolo.marcacao="<>",
     posicao.marcacao=T,
     ignorar.linhas=ignorar,
     stoplist=NULL,
     novos.arquivos="marcacoes")¶
```

A definição acima, no entanto, tem duas limitações: (i) para palavras com mais de uma ocorrência de /r/ em coda (por exemplo, "supermercado"), ela marca apenas uma ocorrência da variável; e (ii) não localiza ocorrências em final de palavra (como "mulher"), que não têm uma consoante após /r/.

A primeira decorre do modo como o R processa os objetos: ao encontrar um item lexical como "supermercado", o R encontra a primeira ocorrência de /r/ em coda na sequência "erm" (uma vogal, o grafema "r" e uma consoante), e a segunda ocorrência "erc" se encaixa na definição de "zero, um ou mais caracteres seguido(s) da fronteira de palavra". Tais caracteres são exauridos na leitura da sequência, o R insere a marcação <> e segue rastreando as demais ocorrências no restante do *corpus*.

Ao invés de tentar formular uma expressão regular mais complexa, que tenha que encontrar casos com uma ou mais ocorrências, pode-se simplesmente criar uma nova que localizará apenas as palavras com duas ocorrências de /r/ em coda, e rodar a função novamente sobre os *arquivos já marcados* com uma ocorrência por palavra. A ocorrência de dois segmentos de /r/ em coda pode ser definida como:

(42)

```
[aáâeéêiíoóôuú]r[bcçdfgjklmnpqstvwxz].*?[aáâeéêiíoóôuú]r
[bcçdfgjklmnpqstvwxz].*?\\b
```

ou seja, a ocorrência de uma vogal, grafema "r" e uma consoante; zero, um ou mais caracteres; outra vogal, outro grafema "r" e outra consoante; zero, um ou mais caracteres; e a fronteira de palavra. Desse modo, podemos apagar os arquivos originais (tendo feito, claro, uma cópia desses em outra pasta segura) e rodar a função abaixo apenas sobre os arquivos "marcacoes".

(43)

```
>identificacao(padrao="[aáâeéêiíoóôuú]r[bcçdfgjklmnpqstvwxz].*?
      [aáâeéêiíoóôuú]r[bcçdfgjklmnpqstvwxz].*?\\b",
      simbolo.marcacao="<>",
      posicao.marcacao=T,
      ignorar.linhas=ignorar,
      stoplist=NULL,
      novos.arquivos="marcacoes")¶
```

Por fim, solução semelhante pode ser aplicada aos casos de /r/ em final de palavra[9]. O padrão pode ser atualizado como em (44) abaixo – uma sequência de vogal, o grafema *r* e fronteira de palavra – e a função pode ser rodada sobre os arquivos agora chamados "marcacoes-marcacoes-...":

(44)

```
>identificacao(padrao="[aáâeéêiíoóôuú]r\\b",
      simbolo.marcacao="<>",
      posicao.marcacao=T,
      ignorar.linhas=ignorar,
      stoplist=NULL,
      novos.arquivos="marcacoes")¶
```

9 Você pode pensar: mas por que não incluir \\b dentro da classe de consoantes, de modo que *final de palavra* seja uma opção à presença de outros grafemas consonantais como *b*, *c*, *ç*, *etc.*? Essa solução não seria possível pelo fato de \\b não ser um caractere e, portanto, não poder integrar uma classe de caracteres.

Os arquivos finais "marcacoes-marcacoes-marcacoes-...", sobre os quais a função foi rodada três vezes, são aqueles que contêm, corretamente, a marcação de todas as ocorrências de (-r) em coda silábica[10].

Exemplo 3: "nós" vs. "a gente"

O emprego da variável do pronome de primeira pessoa do plural exemplifica o caso de variáveis que permitem a codificação automática a partir da função `identificacao()`. No caso dos diminutivos, a codificação de acordo com critérios estabelecidos pelo pesquisador, como Mendes (2012), depende de uma análise contextual mais detalhada, caso a caso (ver nota 8). Para as ocorrências de -r em coda silábica, é necessário ouvir os arquivos de áudio para determinar se o informante realizou o segmento como tepe, retroflexo, apagamento, etc. (como será o caso para variáveis fonéticas).

No caso da variável "nós" *vs.* "a gente", podemos elaborar duas expressões regulares, uma para cada variante, e rodar a função duas vezes, de modo semelhante ao que foi feito com /r/ em coda. O argumento `simbolo.marcacao` pode então ser definido como `"<N>"` para os casos de "nós" e `"<G>"` para os casos de "a gente".

Imagine que o pesquisador decidiu analisar todos os casos de "nós", não apenas no caso nominativo, mas também em outras funções sintáticas. A expressão regular deve então abarcar as ocorrências de "nós", "nos", "conosco", "nosso", "nossa", "nossos" e "nossas".

(45)

```
nós\\b|\\bnos\\b|conosco|noss[oa]s?\\b
```

```
>identificacao(padrao="nós\\b|\\bnos\\b|conosco|noss[oa]s?\\b",
     simbolo.marcacao="<N>",
     posicao.marcacao=T,
     ignorar.linhas=ignorar,
     stoplist=NULL,
     novos.arquivos="marcacoes")¶
```

Em (45), as ocorrências de "nos" são delimitadas tanto no início quanto no final da palavra por `\\b`, para evitar que a função localize ocorrências como "a<u>nos</u>", "me<u>nos</u>", "<u>nos</u>tálgico", etc. Ao examinar as marcações nos arquivos de

10 Casos de /r/ em final de palavra seguidos de vogal, como "po<u>r i</u>sso" e "conta<u>r u</u>ma estória", possivelmente devem ser descartados. No entanto, é preferível ouvir tais ocorrências antes de descartá-las, para se certificar que não há uma pausa entre a palavra com /r/ e a vogal.

transcrição, no entanto, percebemos que várias ocorrências marcadas não se referem ao pronome de primeira pessoa do plural. São casos de "nos" preposição ("em" + "os", por exemplo, linha 72 de FabianaB: "... nos finais de semana à noite") e de "nossa" como interjeição (por exemplo, linha 359 de FabianaB: "nossa está molhado aqui").

Cabe aqui pensar se é o caso de rever nossa definição da expressão regular, ou de criar uma `stoplist`. Contudo, diferentemente dos casos discutidos para /e/ nasal ou para o diminutivo, aqui se trata de coincidências lexicais: se excluirmos as ocorrências de "nossa" e "nos", excluiremos também os **verdadeiros** casos de primeira pessoa do plural. Nessa situação, é necessário remover as ocorrências "a mais" uma a uma, analisando-as em seus contextos. Isso pode ser feito diretamente na interface do RStudio, na aba *Files*, através da janela *Localizar/Substituir* ([CTRL]/[Command]+F). Em *Localizar*, digite "`<N>` " (com um espaço após >) e não digite nada em substituir. Nos casos de marcação indevida, clique em *Replace* para apagar a marcação.

Vejamos agora os casos de "a gente". Aqui, a definição da expressão regular é mais simples, já que o pronome "a gente" não apresenta múltiplas formas como "nós". A expressão `a\\sgente` (sequência de "a", espaço, "gente") encontrará tanto as ocorrências do pronome isolado quanto aquelas em que se amalgama as preposições (por exemplo, "na gente", "pra gente"), contanto que não se especifique uma fronteira de palavra no início. No entanto, essa definição também localizará ocorrências como "muita gente", "pouca gente", "tanta gente" – é o caso de definir uma `stoplist`. Portanto:

(46)

```
>stoplist.G<-c("muita\\sgente","tanta\\sgente","pouca\\sgente")¶
>identificacao(padrao="a\\sgente",
       simbolo.marcacao="<G>",
       posicao.marcacao=T,
       ignorar.linhas=ignorar,
       stoplist=stoplist.G,
       novos.arquivos="marcacoes")¶
```

Ao rodar o código acima sobre os arquivos já marcados com <N>, os arquivos finais "marcacoes-marcacoes-..." conterão todas as ocorrências de "nós" e "a gente", devidamente codificadas para a variável dependente.

4.2 Função extracao()

A função `extracao()` busca um padrão em um conjunto de transcrições e retorna os resultados na forma de uma tabela que contém a ocorrência, o contexto precedente e o seguinte, e sua localização na transcrição. Opcionalmente, a função também retorna a codificação da variável dependente e de variáveis sociais. Assim como na função `identificacao()`, é necessário primeiro especificar como diretório de trabalho a pasta em que se encontram as transcrições, através da função `setwd()` ou através da aba *Files* (seção 2.2).

A descrição de cada argumento se encontra a seguir.

`padrao` — Obrigatório. Definição da sequência de caracteres a ser buscada. Deve ser especificado entre aspas.

`palavras.cont.precedente` — Opcional. Numérico. Número de palavras do contexto precedente a serem extraídas. *Default*=5.

`palavras.ocorrencia` — Opcional. Numérico. Número de palavras a serem extraídas para a coluna de ocorrência. *Default*=1.

`palavras.cont.seguinte` — Opcional. Numérico. Número de palavras do contexto seguinte a serem extraídas. *Default*=5.

`stoplist` — Opcional. Vetor com palavras que devem ignoradas na busca do padrão. *Default*=NULL.

`ignorar.linhas` — Opcional. Vetor com identificação dos participantes, cujas falas devem ser ignoradas na busca do padrão. *Default*=NULL.

`var.dependente` — Opcional. Numérico. Localização do caractere de codificação dentro do termo que contém o padrão buscado. *Default*=NULL.

`loc.variaveis.sociais` — Opcional. Vetor com o número das linhas em que se encontram as informações das variáveis sociais a serem extraídas. *Default*=NULL.

`nomes.colunas.variaveis` — Opcional. Vetor com o nome das variáveis sociais a serem extraídas, na mesma sequência do vetor com sua localização. *Default*=NULL.

`file` — Opcional. Nome do arquivo com dados extraídos. *Default*="DadosExtraidos.txt". O nome deve conter a extensao .txt e ser especificado entre aspas.

O único argumento obrigatório nesta função é a especificação do padrão a ser buscado, na forma de uma expressão regular. A especificação desse argumento pode ser feita de modo idêntico ao da função `identificacao()`. Tomemos como exemplo a expressão regular para os diminutivos, como definida em (36) acima (deixando todos os demais argumentos como *default*).

(47)

```
> extracao("inh[oa]s?\\b")
```

Os resultados se encontram em um arquivo denominado "DadosExtraidos.txt", que se localiza no atual diretório de trabalho. Como se trata de uma tabela, é preferível examinar esses dados em uma planilha, no programa Excel ou Calc. Abra o arquivo em um desses programas[11]. A Figura 5 abaixo mostra as primeiras linhas de resultados:

	A	B	C	D	E	F
1	Contexto.Precedente	Ocorrencia	Contexto.Seguinte	GFs sociais...	Localizacao	
2	eu morava na Vila Prudente de	criancinha	até os três anos em		Paragrafo: 43	
3	(uhum) D1 você conhece seus	vizinhos	por exemplo? S1 ah		Paragrafo: 56	
4	S1 ah eu conheço os	vizinhos	do prédio de elevador assim		Paragrafo: 57	
5	ter um pouco essa relação de	vizinhos	assim aqui você acha?		Paragrafo: 64	
6	mudado assim do que na na	minha	época da adolescência que tinha		Paragrafo: 76	
7	na minha época da adolescência que	tinha	pouca coisa D1 uhum		Paragrafo: 76	
8	D1 uhum S1 tem muitos	barzinhos	aqui né tem ali uma		Paragrafo: 78	
9	tem ali uma rua próxima à	minha	casa S1 onde... montaram		Paragrafo: 78	
10	não faz muito tempo né porque	tinha	claro sempre S1 ali		Paragrafo: 87	
11	você morava na Vila Prudente é	pertinho	também né? S1 é		Paragrafo: 100	
12	ah isso tem com os mais	senhorzinhos	assim né S1 porque...		Paragrafo: 133	
13	assim né S1 porque... os	senhorzinhos	da Mooca são aqueles italianos		Paragrafo: 134	
14	Mooca que é um grupo de	velhinhos	que vai... senhorzinhos Dados		Paragrafo: 136	
15	um grupo de velhinhos que vai...	senhorzinhos	Dados Contextuais [risos-S1]		Paragrafo: 136	
16	Mooca não S1 o que	tinha	anteriormente eram as festas do		Paragrafo: 143	
17	isso que tem/ são mais... os	velhinhos	então será que se reúnem		Paragrafo: 151	
18	D1 uhum S1 perto da	minha	casa tem até um bar		Paragrafo: 157	
19	de rock S1 onde uns	senhorzinhos	têm um grupo de seresta		Paragrafo: 158	
20	estar aqui S1 que eu	tinha	que acordar cedo pra pegar		Paragrafo: 183	
21	metrô Dados Contextuais [barulho de	latinha	se abrindo] D1 uhn		Paragrafo: 184	
22	com quem? S1 com a	minha	mãe e com a minha		Paragrafo: 224	
23	a minha mãe e com a	minha	irmã D1 ah		Paragrafo: 224	
24	vocês brin/ você brincava na rua	tinha	essa tradição de brincar na		Paragrafo: 229	
25	era uma criança muito ativa eu	tinha	diversas atividades no dia... então		Paragrafo: 233	

DadosExtraidos

Figura 5 – Arquivo "DadosExtraidos.txt" para busca do padrão `inh[oa]s?\\b`.

11 Pode-se clicar com o botão direito sobre o arquivo, escolher *Abrir com...* e selecionar um programa de planilhas. Alternativamente, no Excel ou no Calc, clique em *Arquivo* > *Abrir* e selecione o arquivo "DadosExtraidos.txt" (no Excel, é necessário especificar a busca com a opção *Todos os arquivos* para que a janela exiba o arquivo criado, que está em formato .txt). Uma janela se abrirá, pedindo que o usuário especifique o tipo de campo (Delimitado), a linha pela qual se deve iniciar a importação (1) e a origem do arquivo. Verifique se os caracteres aparecem corretamente na *Visualização* da parte inferior e, se necessário, mude a opção em *Origem de arquivo*. Clique em *Avançar*. Na segunda etapa, selecione *Tabulação* como delimitador e clique novamente em *Avançar*. Na terceira etapa, selecione *Geral* para o formato dos dados e clique em *Concluir*.

A tabela contém cinco colunas: *Contexto.Precedente*, *Ocorrencia*, *Contexto.Seguinte*, *GFs sociais*... (que está vazia por não havermos especificado esse argumento na função), e *Localização* (com indicação do parágrafo em que se localiza a ocorrência em sua respectiva transcrição). As palavras-alvo estão separadas na coluna B: veja que, nessa extração, foram localizadas as palavras "vizinhos", "minha", "tinha" etc. – que não são diminutivos – já que não especificamos a `stoplist` em (47) acima. No total, houve 1.706 ocorrências de palavras terminadas em -inh- no *minicorpus* "dmsocio-SP2010".

Como as ocorrências estão separadas, essa coluna pode nos auxiliar a identificar *todas* as ocorrências de palavras terminadas em -inh- que não interessam em nossa análise (lembre-se que, em (38), havíamos especificado as palavras indesejadas que encontramos apenas na entrevista com FabianaB). Tanto o Excel quanto o Calc possuem ferramentas para remover valores duplicados. No Excel, copie a coluna B para uma coluna vazia (preferencialmente, em uma outra planilha); com os dados selecionados, clique em *Dados* > *Remover Duplicatas*. No Calc, o processo envolve alguns passos a mais: copie a coluna B para uma coluna vazia, e com os dados selecionados clique em *Dados* > *Filtros* > *Filtro padrão*. Na janela que se abrir, na primeira linha, escolha *Não vazio* no campo *Valor*. Clique em *Mais opções*, selecione *Sem duplicação* e *Copiar resultados para* e, em seguida, clique em uma célula vazia da tabela. Clique em *OK*.

Das 1.706 ocorrências, há 261 valores únicos. A partir dessa lista – mais prática do que examinar a lista de 1.706 ocorrências, ou mesmo reler todas as transcrições – o pesquisador pode com facilidade criar a lista completa de palavras que devem constar no `stoplist`. Esse caso exemplifica como a função `extracao()` pode ser empregada na *exploração* de dados do *corpus*, sem o objetivo específico de criar uma planilha de codificação.

O pesquisador também pode empregá-la, por exemplo, para testar uma expressão regular, caso não tenha certeza de que a definição formulada localiza corretamente todas e somente as ocorrências referentes à sua variável. Desse modo, ela pode ser utilizada *antes* de aplicar a função `identificacao()`, para verificar rapidamente que tipos de dados se encaixam na expressão regular.

A função `extracao()` também pode ser usada para extrair dados já identificados da variável sociolinguística – finalidade para a qual foi criada de fato – e evitar o trabalho braçal de copiar e colar centenas ou milhares de ocorrências para uma planilha de codificação. Caso o pesquisador já tenha identificado as ocorrências (seja com a função `identificacao()`, seja manualmente) e as codificado no arquivo de transcrições, o padrão a ser buscado pode ser especificado simplesmente com os símbolos usados para marcação de ocorrências (por exemplo, "<N>|<G>").

Para exemplificar esse caso, vamos utilizar os dados do *minicorpus* "dmsocio-codif-R". Trata-se de outro conjunto de 12 entrevistas sociolinguísticas, que

fazem parte do *corpus* de 118 gravações analisadas por Oushiro (2011), já ouvidas e codificadas quanto à pronúncia de (-r) em coda silábica. Os seguintes códigos foram utilizados:

T	Tepe
R	Aproximante retroflexa
A	Apagamento
H	Aspirada (fricativa velar ou glota, surda ou sonora)
V	Realização intermediária ou duvidosa (inclui trechos com muito ruído de fundo)
M	Dados metalinguísticos (por exemplo, "cariocas falam po[x]ta")
E	Palavras estrangeiras (por exemplo, "Big Brother", "Museu d'Orsay")

Há múltiplas possibilidades de análise: podem-se extrair todos os dados, a fim de verificar a sua distribuição e frequências no *corpus*; podem-se também desconsiderar as ocorrências de dados metalinguísticos (que não se referem a realizações "espontâneas" do falante), de palavras estrangeiras (cuja pronúncia pode ser aportuguesada ou não) e de realizações intermediárias (nas quais o pesquisador ficou em dúvida quanto à codificação); pode-se analisar o apagamento de /r/, em contraposição às suas variantes não-apagadas tepe, retroflexo e aspirada; pode-se analisar apenas a pronúncia de tepe *vs*. retroflexo, que são as variantes mais prototípicas da variedade paulistana e se associam mais fortemente às identidades locais.

Essas decisões, como já mencionado, dependem das questões que norteiam a pesquisa. Se nos decidirmos pelo último caso (tepe *vs*. retroflexo), podemos definir o padrão a ser extraído simplesmente como "<T>|<R>" (ou "<[TR]>"). Examinemos, também, os demais argumentos da função `extracao()`.

Palavras.cont.precedente e *palavras.cont.seguinte* são argumentos numéricos que especificam quantas palavras devem ser extraídas antes e depois da ocorrência. Em geral, para variáveis fonéticas, poucas palavras (digamos, 3 ou 4) bastam para estabelecer um contexto que permita a codificação de outras variáveis linguísticas, como contexto fônico precedente, contexto fônico seguinte, classe morfológica, etc. Para variáveis morfológicas, sintáticas e discursivas, normalmente é preferível a extração de um contexto textual maior (digamos, 15 ou 20 palavras). Para -r em coda, vamos estabelecer 4 palavras.

O número de palavras da coluna *Ocorrencia* pode conter apenas o padrão buscado (neste caso, "<T>" ou "<R>"*)*, ou incluir um número maior de palavras

antecedentes. No presente exemplo, é interessante extrair a codificação da ocorrência junto com o dado em si (a palavra com /r/ em coda), de modo que vamos estabelecer `palavras.ocorrencia=2`.

Os argumentos *stoplist* e *ignorar.linhas* funcionam de forma idêntica à sua aplicação na função `identificacao()`. Neste caso, ambos podem ser =*NULL*: não há palavras que devem ser evitadas e as marcações de ocorrência foram feitas apenas na fala do informante.

O argumento *var.dependente* indica, numericamente, a localização do caractere de codificação da variável dependente dentro do *padrão* estabelecido. Neste exemplo, o código da variante empregada pelo falante em cada ocorrência se encontra na segunda posição de `<T>` e `<R>`; estabelecemos, portanto, `var.dependente=2`.

O argumento *loc.variaveis.sociais* pressupõe a existência de um cabeçalho em cada transcrição. Ele é preenchido com um vetor que indica a localização de informações da ficha social do informante (por exemplo, sexo/gênero, faixa etária, etc.) de acordo com a linha em que aparecem. O cabeçalho das transcrições do *minicorpus* "dmsocio-codif-R" indica as seguintes informações:

`#cab`	
`2009`	Ano de gravação
`F`	Sexo: F – feminino; M - masculino
`27`	Idade em anos
`1`	Faixa etária: 1 – 20 a 34; 2 – 35 a 59; 3 – 60 ou mais
`C`	Escolaridade: C – até Ens. Médio; S – Ens. Superior
`P`	Região de residência: C – bairro central; P – bairro periférico
`L`	Zona de residência: N – norte; S – sul; L – leste; O – oeste; C – centro
`0`	Geração na cidade: 0 – pais não paulistanos; 1 – mãe ou pai paulistano; 2 – mãe e pai paulistano; 3 – um(a) avô/avó paulistano; 4 – dois ou mais avós paulistanos
`I`	Origem dos pais: P – São Paulo-capital; I – interior SP/MG; N – região N/NE; E – estrangeira; X – mista
`Z`	Mobilidade: B – sempre morou no mesmo bairro; Z – sempre morou na mesma zona; M – já morou em diferentes zonas ou outra cidade
`PamelaR`	Pseudônimo do informante
`Artur Alvim`	Bairro de residência
`D1: Livia Oushiro`	Documentador
`Duração total: 01h10min32seg`	
`Início: 00h00min00seg`	
`Fim: 01h00min00seg`	
`Comentários:`	

O cabeçalho, é claro, pode diferir de *corpus* para *corpus*. Se foi digitado de modo consistente em todas as transcrições, é possível indicar em um vetor a localização das variáveis sociais relevantes que o pesquisador deseja incluir em sua análise. O vetor `var.sociais` abaixo indica a localização das variáveis sexo/gênero, faixa etária, escolaridade, região de residência e pseudônimo do informante que se encontram, respectivamente, nas linhas 3, 5, 6, 7 e 12 dos cabeçalhos.

(48)

```
> var.sociais<-c(3, 5, 6, 7, 12) ¶
```

O argumento *nomes.colunas.variaveis*, por sua vez, especifica o nome que deve ser atribuído às variáveis sociais, na mesma ordem em que foram especificadas para *var.sociais*.[12] Neste caso:

(49)

```
> nomes.var.sociais<-c("sexo.genero", "faixa.etaria",
          "escolaridade", "regiao.residencia",
          "informante")¶
```

Por fim, o argumento *file* especifica o nome do arquivo com a tabela de dados extraídos. O *default* "DadosExtraidos.txt" pode ser modificado pelo usuário, atentando-se ao requisito de que o nome deve vir entre aspas e com a extensão .txt. Aqui, usaremos o nome "DadosExtraidos-RT.txt".

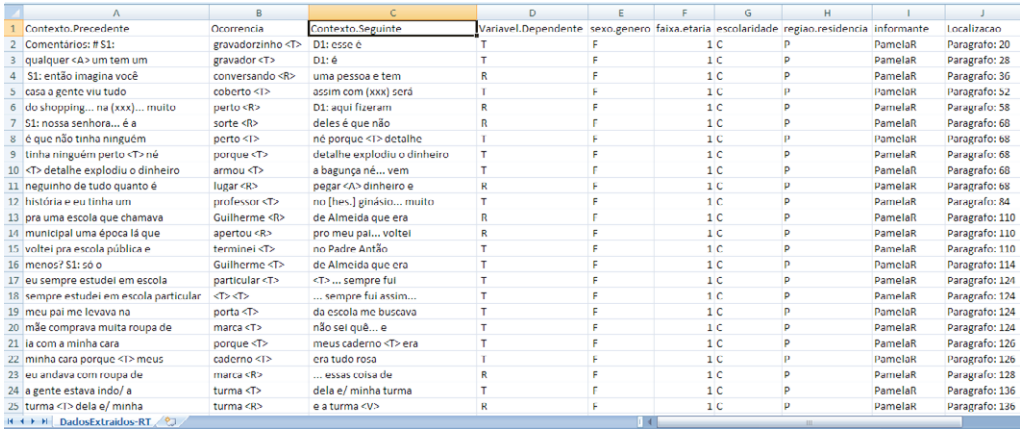
Desse modo, temos então:

(50)

```
> extracao(padrao="<T>|<R>",
     palavras.cont.precedente=4,
     palavras.ocorrencia=2,
     palavras.cont.seguinte=4,
     stoplist=NULL,
     ignorar.linhas=NULL,
     var.dependente=2,
     loc.variaveis.sociais=var.sociais,
     nomes.colunas.variaveis=nomes.var.sociais,
     file="DadosExtraidos-RT.txt")
```

12 É possível especificar somente um vetor para `loc.variaveis.sociais` e deixar `nomes.colunas.variaveis` como =NULL, caso em que as variáveis serão nomeadas "Var.1", "Var.2", etc. No entanto, se os nomes das variáveis forem especificados, não é possível deixar o argumento `loc.variaveis.sociais` como =NULL, já que o script demandará a localização das variáveis.

O resultado é uma tabela, cujas primeiras linhas podem ser visualizadas na Figura 6. Em alguns minutos, o pesquisador tem em mão uma tabela de dados semipronta para ser analisada em programas como GoldVarb X ou RBrul, algo que poderia levar horas ou dias se fosse criada manualmente.



	A	B	C	D	E	F	G	H	I	J
1	Contexto.Precedente	Ocorrencia	Contexto.Seguinte	Variavel.Dependente	sexo.genero	faixa.etaria	escolaridade	regiao.residencia	informante	Localizacao
2	Comentários: # S1:	gravadorzinho <T>	D1: esse é	T	F	1	C	P	PamelaR	Paragrafo: 20
3	qualquer <A> um tem um	gravador <T>	D1: é	T	F	1	C	P	PamelaR	Paragrafo: 28
4	S1: então imagina você	conversando <R>	uma pessoa e tem	R	F	1	C	P	PamelaR	Paragrafo: 36
5	casa a gente viu tudo	coberto <I>	assim com (xxx) será	I	F	1	C	P	PamelaR	Paragrafo: 52
6	do shopping... na (xxx)... muito	perto <R>	D1: aqui fizeram	R	F	1	C	P	PamelaR	Paragrafo: 58
7	S1: nossa senhora... é a	sorte <R>	deles é que não	R	F	1	C	P	PamelaR	Paragrafo: 68
8	é que não tinha ninguém	perto <I>	né porque <I> detalhe	I	F	1	C	P	PamelaR	Paragrafo: 68
9	tinha ninguém perto <T> né	porque <T>	detalhe explodiu o dinheiro	T	F	1	C	P	PamelaR	Paragrafo: 68
10	<T> detalhe explodiu o dinheiro	armou <T>	a bagunça né... vem	T	F	1	C	P	PamelaR	Paragrafo: 68
11	neguinho de tudo quanto é	lugar <R>	pegar <A> dinheiro e	R	F	1	C	P	PamelaR	Paragrafo: 68
12	história e eu tinha um	professor <T>	no [hes.] ginásio... muito	T	F	1	C	P	PamelaR	Paragrafo: 84
13	pra uma escola que chamava	Guilherme <R>	de Almeida que era	R	F	1	C	P	PamelaR	Paragrafo: 110
14	municipal uma época lá que	apertou <R>	pro meu pai... voltei	R	F	1	C	P	PamelaR	Paragrafo: 110
15	voltei pra escola pública e	terminei <T>	no Padre Antão	T	F	1	C	P	PamelaR	Paragrafo: 110
16	menos? S1: só o	Guilherme <T>	de Almeida que era	T	F	1	C	P	PamelaR	Paragrafo: 114
17	eu sempre estudei em escola	particular <T>	<T> ... sempre fui	T	F	1	C	P	PamelaR	Paragrafo: 124
18	sempre estudei em escola particular	<T> <T>	... sempre fui assim...	T	F	1	C	P	PamelaR	Paragrafo: 124
19	meu pai me levava na	porta <T>	da escola me buscava	T	F	1	C	P	PamelaR	Paragrafo: 124
20	mãe comprava muita roupa de	marca <T>	não sei quê... e	T	F	1	C	P	PamelaR	Paragrafo: 124
21	ia com a minha cara	porque <T>	meus caderno <T> era	T	F	1	C	P	PamelaR	Paragrafo: 126
22	minha cara porque <I> meus	caderno <I>	era tudo rosa	I	F	1	C	P	PamelaR	Paragrafo: 126
23	eu andava com roupa de	marca <R>	... essas coisa de	R	F	1	C	P	PamelaR	Paragrafo: 128
24	a gente estava indo/ a	turma <T>	dela e/ minha turma	T	F	1	C	P	PamelaR	Paragrafo: 136
25	turma <I> dela e/ minha	turma <R>	e a turma <V>	R	F	1	C	P	PamelaR	Paragrafo: 136

DadosExtraidos-RT

Figura 6 – Dados de tepe e retroflexo extraídos da amostra "dmsocio-codif-R".

4.3 Função `amostragem()`

A função `amostragem()` seleciona aleatoriamente um determinado número de dados, com base em uma coluna de referência – por exemplo, 50 dados por informante. A tabela "DadosExtraidos-RT.txt" revela que houve um total de 3.282 dados de /r/ tepe ou retroflexo na pequena amostra de 12 entrevistas sociolinguísticas – uma média de 274 dados por falante. Tal costuma ser o caso para variáveis fonéticas, em que não raro se obtêm milhares ou dezenas de milhares de ocorrências da variável sociolinguística. De fato, em sua amostra de 118 informantes paulistanos, Oushiro (2012) obteve mais de 70 mil dados de /r/ em coda silábica e cerca de 35 mil para as variantes tepe e retroflexa.

Em estudos sociolinguísticos de variáveis fonéticas, nem sempre é necessário trabalhar com todos os dados (WOLFRAM, 1993). De um lado prático, uma subamostra torna a tarefa de codificar as variáveis independentes (em nosso caso, apenas as linguísticas, já que as variáveis sociais já estão codificadas) mais manejável. Mais importante, do lado teórico-metodológico, uma subamostra – contanto que aleatória – costuma revelar os mesmos padrões de variação e permite controlar o peso que cada informante tem nos resultados finais, se cada um

pelo mesmo número de dados. Ademais, é preferível analisar um menor número de dados de uma maior quantidade de informantes, do que analisar todos de poucos informantes – a amostra com maior quantidade de informantes tenderá a revelar padrões que mais se aproximam daqueles empregados na comunidade.

É importante frisar, no entanto, que a aplicação da função `amostragem()` é recomendada apenas em casos de variáveis sociolinguísticas muito frequentes, cuja distribuição é razoavelmente balanceada na amostra. Em outros casos, é preferível analisar o conjunto completo de dados.

Essa função requer a instalação de um pacote adicional, chamado NCStats (OGLE, s/d), que não consta na instalação base do R. A própria função `amostragem()` verifica se o pacote já está instalado no computador e, em caso negativo – como provavelmente será na primeira vez que a função for empregada –, busca-o na Internet e faz sua instalação. Se o computador não estiver conectado à internet, aparecerá uma mensagem de erro avisando que não existe o pacote NCStats. A conexão à internet, portanto, é requerida, mas apenas na primeira utilização da função. Nas próximas vezes, o programa já estará instalado.

A função contém quatro argumentos, descritos e exemplificados a seguir:

`numero.dados` Opcional. Numérico. Número de dados a serem selecionados aleatoriamente da coluna de referência. *Default* = 50.

`coluna.referencia` Obrigatório. Numérico. Coluna que deve ser usada como referência para amostragem. A = 1, B = 2, C = 3, etc.

`data` Obrigatório. *Dataframe* com planilha de dados a serem amostrados.

`novo.arquivo` Opcional. Nome do novo arquivo com os dados amostrados. *Default*="DadosAmostrados.txt"

Suponha que, do conjunto total de 3.282 dados de tepes e retroflexos, da amostra "dmsocio-codif-R", o pesquisador decida extrair 100 dados aleatórios por falante, identificados na coluna I (=coluna 9 da esquerda para a direita) da tabela "DadosExtraidos-RT". Primeiro, devem-se carregar os dados em um *dataframe*, através da função `read.table()`. Se se trata do mesmo arquivo gerado com a função `extracao()`, seus argumentos serão:

(51)

```
>dadosR<-read.table(file=choose.files(), header=T, sep="\t",
quote="", comment.char="")¶
```

A função `amostragem()` pode então ser definida como

(52)

```
> amostragem(numero.dados=100,
      coluna.referencia=9,
      data=dadosR,
      novo.arquivo="DadosAmostrados-RT.txt")¶
```

O arquivo "DadosAmostrados-RT.txt", com 1.200 dados (100 dados x 12 falantes) é criado no atual diretório de trabalho.

CONSIDERAÇÕES FINAIS

Este texto é uma introdução prática ao uso do programa R como ferramenta para análises sociolinguísticas. Além de apresentar alguns conceitos e funções básicas, demonstrou-se a aplicação de três funções desenvolvidas especificamente para certas tarefas de análise sociolinguística, que compreendem a preparação do arquivo de dados: `identificacao()`, `extracao()` e `amostragem()`. Os dois primeiros, em especial, também podem ser empregados como métodos de exploração de dados do *corpus*, na fase de análise qualitativa.

O principal objetivo dessas funções é automatizar a realização de tarefas mecânicas e repetitivas, que não constituem a verdadeira tarefa do sociolinguista; ao reduzir o tempo empregado nessas tarefas, o pesquisador pode se dedicar àquilo que de fato constitui o seu papel: à descrição e à análise da heterogeneidade ordenada. O pesquisador terá mais tempo para realizar análises estatísticas, interpretar os resultados e refinar suas análises.

O programa R também pode ser utilizado na realização de diferentes tipos de análises estatísticas, tópico que está além do escopo deste artigo. No entanto, espera-se que, ao se familiarizar com o programa e com as possibilidades de tratamento de dados que ele oferece, o leitor se encoraje a aprofundar seus conhecimentos sobre o R. Ao final do artigo, encontra-se uma pequena lista de referências selecionadas para esse fim.

Boas análises!

LEITURAS RECOMENDADAS

- Para saber mais sobre cada função, digite "?nomedafunção" no console do R. Por exemplo, ?scan. A descrição da função, de seus argumentos e de seus usos aparece na aba *Help* do RStudio;

- GRIES, S. Th. *Quantitative Corpus Linguistics with R*. A practical introduction. New York/London: Routledge, 2009a.

  Trata-se de um livro prático e didático sobre o uso do R para processamento de dados textuais. A obra se volta, principalmente, para a Linguística de *Corpus*, mas exemplifica muitos usos que são de grande interesse para a Sociolinguística. O pacote dmsocio foi desenvolvido com base em seu conteúdo.
- BAAYEN, R. H. *Analyzing Linguistic Data*: a practical introduction to statistics using R. Cambridge: Cambridge University Press, 2008;
- DALGAARD, P. *Introductory statistics with R*. New York: Springer, 2008;
- GRIES, S. Th. *Statistics for Linguistics with R*. Berlin/New York: Mouton de Gruyter, 2009b.

  Após a preparação do arquivo de dados, o próximo passo – e mais importante – é a análise propriamente dita. Essas obras introduzem o leitor a diversos tipos de análises estatísticas, para além dos programas GoldVarb X e RBrul. Dalgaard (2008) é para um público geral e explica diversos conceitos básicos de estatística. Baayen (2008) e Gries (2009b) se voltam especificamente ao público de linguistas.

AGRADECIMENTO

Agradeço ao Grupo de Estudos do R da USP, sobretudo a Fernanda Canever, pelos constantes diálogos em nossa aprendizagem sobre o programa e por testar as funções em outro sistema operacional.

REFERÊNCIAS

BAAYEN, R. H. *Analyzing linguistic data*: a practical introduction to statistics using R. Cambridge: Cambridge University Press, 2008.

BATTISTI, E. *Elevação das vogais médias pretônicas em sílaba inicial de vocábulo na fala gaúcha*. Porto Alegre, 1993. 125f. Dissertação (Mestrado). Universidade Federal do Rio Grande do Sul.

_____. A redução dos ditongos nasais átonos. In: BISOL, L.; BRESCANCINI, C. (eds.), *Fonologia e variação: recortes do português brasileiro*. Porto Alegre: EdiPUCRS, 2002.

BAYLEY, R. The quantitative paradigm. In: CHAMBERS, J.K.; TRUDGILL, P.; SCHILLING-ESTES, N. (eds.), *The Handbook of Language Variation and Change*, p. 117-141. Malden, MA: Blackwell, 2002.

CASTILHO, A.; PRETI, D. (eds.) *A linguagem falada culta na cidade de São Paulo*: materiais para seu estudo, vol. I – Elocuções Formais. São Paulo: T.A. Queiroz, 1986.

CELIA, G. F. *As vogais médias pretônicas na fala culta de Nova Venécia*. Campinas, 2004. 114f. Dissertação (Mestrado). IEL/Unicamp.

CUNHA, C.; CINTRA, L. *A Nova gramática do português contemporâneo*. 3ª edição revista. Rio de Janeiro: Lexikon Informática, 2007.

DALGAARD, P. *Introductory statistics with R*. New York: Springer, 2008.

GRIES, S. Th. *Quantitative corpus linguistics with R*: a practical introduction. New York/ London: Routledge, 2009a.

_____. *Statistics for linguistics with R*. Berlin/New York: Mouton de Gruyter, 2009b.

GUY, G. R. The quantitative analysis of linguistic variation. In: PRESTON, D. (ed.), *American Dialect Research*, p. 223-249. Amsterdam: Benjamins, 1993.

_____. Linguistic variation in brazilian portuguese: aspects of the phonology, syntax and language history.Pennsylvania, 1981. 406f. Tese (Doutorado). University of Pennsylvania.

LABOV, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language*, vol. 45(4): 715-762, 1969.

MELLO, H.; RASO, T. Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. *Revista Portuguesa de Humanidades*, Portugal, v.13, n. 1, p. 301-325, 2009.

MENDES, R. B. Gênero/sexo, variação linguística e intolerância. In: BARROS, D. L. P. (ed.): *Preconceito e intolerância:* Reflexões linguístico-discursivas. São Paulo: Editora do Mackenzie, p. 171-192, 2011.

_____. Diminutivos como marcadores de sexo/gênero. *Revista Linguística*, *v*.8, n.1, p.113-124, 2012.

MENDES, R. B.; OUSHIRO, L. Documentação do Projeto SP2010 – Construção de uma amostra da fala paulistana, 2013. Disponível em <http://projetosp2010.fflch.usp.br/producao-bibliografica>. Acessado em: 01 maio 2014.

OGLE, D. H. NCStats package, v. 0.4.0. Disponível em <https://rforge.net/NCStats/>. acessado em: 01 maio 2014.

OUSHIRO, L. Relatório científico parcial apresentado à FAPESP. (Projeto: Identidade na pluralidade: produção e percepção linguística na cidade de São Paulo, Processo no. 2011/09122-6), 2012.

_____. Ditongação do /e/ nasal no português paulistano. In: SEMINÁRIO DO GEL, 61., 2013, São Paulo. Programação – 61º Seminário do GEL; 2013. v. 1.

OUSHIRO, L.; MENDES, R. B. A pronúncia de /r/ em coda silábica no português paulistano. *Revista do* GEL, São Paulo, v. 8, n. 2, p. 66-95, 2013.

PAIVA, M. C.; SCHERRE, M. M. P. Retrospectiva sociolinguística: contribuições do PEUL. DELTA [*online*], vol.15, n.spe, p. 201-232, 1999.

R CORE TEAM (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <http://www.R-project.org/>. Acessado em: 01 maio 2014.

SCHWINDT, L. C.; SILVA, T. B. da. Panorama da redução da nasalidade em ditongos átonos finais do português do sul do Brasil. In: BISOL, L.; COLLISCHONN, G. (eds.). *Português do Sul do Brasil*: variação fonológica, Porto Alegre: EdiPUCRS, p. 13-33, 2009.

TENANI, L.; GONÇALVES, S. C. L. *Manual do sistema de transcrição de dados* (v.5) – Projeto ALIP (Amostra Linguística do Interior Paulista). Ms, s/d.

TENANI, L.; SILVEIRA, A. A. M. O alçamento das vogais médias na variedade culta do noroeste paulista. *Alfa: Revista de Linguística*, São Paulo: v. 52, n. 2, p.447-464, 2008.

VIEGAS, M. C. (1987). Alçamento das vogais médias pretônicas: uma abordagem sociolinguística. Minas Gerais, 1987. Dissertação (Mestrado). Universidade Federal de Minas Gerais.

VOTRE, S. J. *Aspectos da variação fonológica na fala do Rio de Janeiro*. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 1978.

WEINREICH, U.; LABOV, W.; HERZOG, M. I. Empirical foundations for a theory of language change. In: LEHMANN, W.P.; MALKIEL, Y. (eds.). *Directions for Historical Linguistics*: A Symposium. Austin: University of Texas Press, 1968.

WOLFRAM, W. Identifying and interpreting variables. In: PRESTON, D. R. (ed.), *American Dialect Research*, Amsterdam/Philadelphia: John Benjamins, p. 193-221, 1993.

ZILLES, A. The development of a new pronoun: The linguistic and social embedding of a gente in Brazilian Portuguese. *Language Variation and Change*, vol. 17, p. 19-53, 2005.