Doi: http://dx.doi.org/10.5007/1984-8412.2017v14n2p2173

A FORMAÇÃO DE UM BANCO DE DADOS DE FALA DE TERESINA (PI): UM ESTUDO DE CASO

LA FORMACIÓN DE UNA BASE DE DATOS DE HABLA DE TERESINA (PI): UN ESTUDIO DE CASO

THE FORMATION OF A SPEECH DATABASE FROM TERESINA (PI): A CASE STUDY

Francisca da Cruz Rodrigues Pessoa[*]
Jânia Martins Ramos[**]
Universidade Federal de Minas Gerais

RESUMO: Este artigo enumera iniciativas de registro de dados de fala representativos da comunidade residente em Teresina (PI). Seus objetivos são descrever o estágio atual da construção dos diferentes *corpora*, compará-los, e sugerir que seja organizado um banco de dados único. Para encaminhamento dessa proposta, argumenta-se a favor (i) da adoção de uma política de autoria para a edição de *corpora*, e (ii) da reunião de amostras de entrevistas coletadas por pesquisadores que atuam em projetos independentes, quer sejam sociolinguistas quer sejam de áreas afins. A opção pela análise de caso, por tomar como ponto de partida uma situação real, oportuniza colocar em foco uma situação que certamente não é excepcional nem é estranha a numerosos pesquisadores brasileiros.
PALAVRAS-CHAVE: Banco de dados. C*orpus*. Teresina. Fala.

RESUMEN: Este artículo enumera iniciativas de registro de datos de habla representativos de la comunidad residente en Teresina (PI). Sus objetivos son describir la evolución actual de la construcción de los diferentes corpus, cotejarlos, y proponer que sea organizada una base de datos única. Para afirmar esa propuesta, se argumenta a favor (i) de la adopción de una política de autoría para la edición de corpus, y (ii) de la reunión de muestras de entrevistas recolectadas por investigadores que actúan en proyectos independientes, sean sociolingüistas sean de áreas relacionadas. La decisión por el análisis de caso, teniendo como punto de partida

[*]Doutora em Estudos Linguísticos – UFMG | Docente da Secretaria de Educação do Estado do Piauí (SEDUC/PI). E-mail: franrodriguespessoa@hotmail.com.

[**]Professora Doutora do Quadro Permanente do Programa de Pós-Graduação em Estudos Linguísticos da UFMG e Bolsista de Produtividade em Pesquisa do CNPq. E-mail: jania.ramos@gmail.com.

una situación real, permite poner en foco una situación que seguramente no es excepcional ni es extraña para numerosos investigadores brasileños.
PALABRAS CLAVE: Base de datos. *Corpus*. Teresina. Habla.

ABSTRACT: This article lists initiatives of recording of speech data typical of the community resident in Teresina (PI). Its objectives are to describe the current development stage of different *corpora*, to compare them, and to suggest an organization of a single database. For submitting this proposal, it is argued in favor of (i) the adoption of an authorship policy for *corpora* editing, and of (ii) the assembling of interview samples gathered by researchers involved in independent projects, be they sociolinguists or from related areas. Opting for a case analysis, once it depicts a real situation as its starting point, provides the opportunity to shed light on a situation that, certainly, is neither exceptional nor unusual to a large number of Brazilian researchers.
KEYWORDS: Database. *Corpus*. Teresina. Speech.

1 INTRODUÇÃO

O interesse pela língua falada no Brasil foi impulsionado pelo Projeto de Estudo da Norma Linguística Urbana (Projeto NURC)[1] e, a partir de então, dezenas de novos projetos foram empreendidos. Em geral, cada projeto produziu um novo *corpus* ou ampliou um *corpus* já formado. Quase sempre esses *corpora* foram organizados em conformidade com as diretrizes metodológicas da Sociolinguística Variacionista (LABOV, 2008 [1972]; TAGLIAMONTE, 2006)[2].

Por serem numerosos e metodologicamente similares, seria de grande interesse para os estudos sociolinguísticos no Brasil se os pesquisadores pudessem ter reunidos, num único *site*, os *links* de todos os *corpora*, quer livres quer de uso restrito. Isso representaria um grande ganho no que diz respeito à economia de recursos humanos e financeiros. Embora reconhecidamente relevante, esse empreendimento não tem sido, até o momento, definido como meta pelas entidades representativas da área, como o Grupo de Trabalho de Sociolinguística da Associação Nacional de Programas de Pós-Graduação em Letras e Linguística (ANPOLL) ou pela Associação Brasileira de Linguística (ABRALIN). Algumas iniciativas, entretanto, já podem ser registradas, como, por exemplo, a realização de um simpósio sobre *Corpora*, no Seminário do Grupo de Estudos Linguísticos de São Paulo (GEL) em 2013[3], e a recolha de notícias sobre corpora, por parte da Coordenação do Grupo de Trabalho (GT) de Sociolinguística da ANPOLL, no biênio 2014-2015.

Mas como viabilizar uma proposta de centralização de informação sobre *corpora*? Que tipo de informação seria viável? Neste artigo, vamos apresentar uma proposta. Tomaremos como exemplo as iniciativas adotadas para o registro da língua falada de moradores da cidade de Teresina (PI). Optamos por apresentar um estudo de caso, bem pontual, sobre um dialeto ainda pouco conhecido, na medida em que poucos estudos o tomaram como objeto de análise. É nosso objetivo apresentar e discutir um formato de maximização das potencialidades de uso dos produtos de trabalhos realizados até o momento.

2 FALA DOS TERESINENSES

[1] Ver Castilho (1990).

[2] Conforme se lê em Milroy e Gordon (2003, p.23-48), é central nesta abordagem a suposição de que a variação linguística é social e gramaticalmente motivada. Uma consequência é a exigência de que o investigador obtenha tipos e quantidades de dados linguísticos suficientes, socialmente contextualizados. É necessário, portanto, compor amostras representativas da língua de grupos de falantes, o que implica a necessidade de considerar uma ampla gama de sujeitos. Amostras linguísticas, segundo Romaine (1980), são, de modo geral, muito pequenas para assegurar representatividade do ponto de vista estatístico, A saída tem sido formar amostras organizadas a partir de um perfil previamente definido pelo pesquisador, o que vai exigir a explicitação dos critérios adotados e a demonstração de que tais critérios são claros e bem motivados. Para coleta de dados são usados questionários escritos, pesquisas rápidas e anônimas, entrevistas sociolinguísticas, observação participativa ou, ainda, uma interação entre estas estratégias de pesquisa de campo.

[3] Simpósio cujos resultados foram publicados em formato de livro (FREITAG, 2014).

Como é o português falado em Teresina (PI)? O que dizer do dialeto de um centro urbano regional cuja área de influência inclui parte do próprio estado e uma pequena área do Maranhão? Situada dentro da Região de influência de Fortaleza, Teresina concentra atividades administrativas, dispõe de serviços de ensino e saúde, comércio e lazer (IBGE, 2007). Sua zona urbana atual possui de 242 km e teve início há 166 anos. A evolução populacional vai de 21.691 (Censo de 1872) a 814.439 (Censo de 2010), motivada por fatores de ordem natural, as secas. A partir dos anos setenta do século XX sua população quadruplicou; cerca de 70% tem renda igual ou inferior a 2 salários mínimos (Censo de 2010). Esse "[...] centro de migração, principalmente de famílias de baixa renda, provenientes do interior do Estado" (LIMA, 2011, p. 77). Teresina é, atualmente, a principal cidade do Estado e se tornou um importante polo de serviços para o meio norte do país, principalmente na área de saúde e indústria têxtil[4].

Mas como é a fala teresinense? É restrita a informação sociolinguística que se tem sobre esse dialeto. Encontramos duas teses de doutorado, dois artigos e quatro trabalhos publicados em Anais. Pesquisando esse material, pudemos obter informações sobre cinco corpora[5]. O primeiro compõe-se de 96 entrevistas; o segundo e o terceiro de 36 entrevistas cada, o quarto de 12 entrevistas e o quinto de 11, formando um total de 191 entrevistas, gravadas no período de 2000 a 2015. Esse resultado nos levou à formulação de novas perguntas: como esses corpora têm sido referidos? Como estão organizados? Foram digitalizados? Estão disponíveis? Na seção a seguir buscaremos responder a essas questões.

3 CASO A CASO

Consideremos inicialmente o *corpus* organizado pelas professoras Maria Auxiliadora Ferreira Lima (UFPI) e Maria Anecy Calland Marques Serra (UFPI). O título do projeto que gerou esse *corpus* denomina-se *Aspectos Gramaticais do Português Falado por Estudantes Teresinenses* (2000-2003). A amostra compõe-se de 96 entrevistas, de 60 minutos de gravação, em média. Os informantes têm idade entre 9 a 20 anos e são moradores do município de Teresina. Nesse *corpus*, cada célula é de 4 informantes, alunos da quinta e nona séries do Ensino Fundamental, e do terceiro ano do Ensino Médio, respectivamente. Outros parâmetros de composição da amostra são o sexo, o poder aquisitivo – esse último depreendido do tipo de escola (pública ou particular) – , profissão e nível de escolaridade dos pais, localização das escolas (centro ou periferia). Estas informações se encontram numa ficha, que foi preenchida, antes da entrevista, pelos informantes. Os temas das entrevistas são esporte, diversão, leitura, política, viagem, cinema, programas de televisão, juventude, droga, violência, etc. As entrevistas foram gravadas em fita cassete e transcritas com ortografia usual, de acordo com normas de transcrição previamente definidas (LIMA; SERRA, 2010.).

Uma subparte desse *corpus* foi divulgada no formato de livro impresso em papel. O acesso ao *corpus* completo é restrito. Alunos de graduação e pós-graduação da Universidade Federal do Piauí têm realizado pesquisas com base nesse *corpus*: menções explícitas ao *corpus* foram encontradas em 1 tese em andamento, 13 dissertações e 03 artigos, embora nem todos esses trabalhos sejam estudos sociolinguísticos.

O total das entrevistas encontra-se digitalizado, tendo sido utilizado o equipamento Plus Deck 2C, no *Laboratório de Estudos em Variação Linguística*, da Faculdade de Letras, da Universidade Federal de Minas Gerais)[6]. Esse *corpus* é citado como: LIMA, Maria

[4] Censo (IBGE, 2010).

[5] De outras cidades do Piauí identificamos dois corpora. O primeiro adota metodologia sociolinguística de natureza etnográfica (GUMPERZ; HYMES, 1972), e a amostra compõe-se de seis falantes adultos da cidade de Angical do Piauí (PI), (SOUSA, 2012). O segundo corpus tem como amostra oito informantes da cidade de Ipiranga (PI), sendo "a maioria mulheres de zona rural analfabetas, de idades entre 45 a 80 anos" (DANTAS; CARVALHO; COSTA, 2014, p 294).

[6] A digitalização foi feita com a colaboração de Francisca da Cruz Rodrigues Pessoa, doutoranda do Programa de Pós-Graduação em Estudos Linguísticos, da Universidade Federal de Minas Gerais, apoiada pela Faculdade Maranhense São José dos Cocais (MA).

Auxiliadora Ferreira; SERRA, Maria Anecy Calland Marques. *Português falado por estudantes teresinenses*. Vol. 1, Teresina: EDUFPI, 2010. A disponibilização do áudio continua restrita.

O segundo *corpus* foi organizado por Ailma do Nascimento Silva e utilizado como objeto de análise em sua tese de doutorado (SILVA, 2009). Compõe-se de entrevistas orais de 36 informantes, sendo 18 do sexo feminino e 18 do sexo masculino, distribuídos em três níveis de escolaridade (Ensino Fundamental, Ensino Médio e Ensino Superior) e três faixas etárias (20-35, 36-50 e +50 anos). Na seleção dos informantes, os critérios adotados foram: ser nascido em Teresina (ou no interior do Estado e chegando na capital antes de 12 anos idade); ter morado na cidade pelo menos 2/3 de sua vida; não ter morado fora do Estado por mais de um ano no período de aquisição da língua nativa (2 a 12 anos). Para a coleta de dados, utilizou-se a técnica da entrevista de experiência pessoal gravada, seguindo a metodologia empregada usualmente em estudos de caráter variacionista. As entrevistas, com média de duração de 45 minutos, foram realizadas entre dezembro de 2006 e março de 2007.

Em trabalhos posteriores de Ailma Silva , o *corpus* não é citado, embora tenha sido utilizado: não recebe um título, e, consequentemente, não é identificado como uma produção científica. A disponibilização do *corpus* é restrita, tendo sido utilizado por orientandos em trabalhos em co-autoria com a organizadora[7].

O terceiro *corpus* foi organizado por Lucirene da Silva Carvalho, tendo sido objeto de análise em sua tese de doutorado (CARVALHO, 2008). A amostra utilizada nesta pesquisa restringiu-se a 36 informantes, sendo 20 da capital e 16 da região norte do estado. Esses 16 informantes, apesar de residirem na capital, estão sempre indo às suas cidades de origem, visto terem, a maioria deles, parentes consanguíneos. Os informantes são provenientes de: Altos (2), Miguel Alves (1), Campo Maior (2), Piripiri (1), Castelo do Piauí (2), Cocal (1), Buriti dos Montes (1), Barras (1), José de Freitas (2), Boa Hora (1), Esperantina (2) e Parnaíba (2).

Imagem 1: Mapa do Estado do Piauí

Fonte: Carvalho (2008, p. 48, [Mapa 2].

[7] Nunes e Silva (2011, p.67) mencionam um "[...] banco de dados de fala coletados a partir de entrevistas pessoais com 36 falantes teresinenses estratificados socialmente por gênero (masculino e feminino), faixa etária (20-35; 36-50 e + 50 anos) e escolaridade (Ensino Fundamental, Ensino Médio e Ensino Superior)".

Os informantes foram distribuídos em três faixas etárias, sendo a primeira de 15-25, a segunda de 26-47, e a terceira, de 47 anos ou mais. A coleta de dados ocorreu entre dezembro de 2006 a abril de 2007. A duração de cada entrevista, em média, é de 30 minutos, e está dividida em dois momentos: a) leitura dirigida, em que se usaram palavras e períodos longos; b) conversa livre, formada de narrativas de experiências pessoais. Cada informante preencheu uma ficha de identificação e assinou um termo de consentimento livre e esclarecido.

O equipamento usado para gravação digital foi Olympus VN -120PC. As entrevistas foram armazenadas em computador e transcritas[8]. A disponibilização para pesquisa é também restrita. Nenhuma citação formal desse *corpus* aparece em outros textos que dele se serviram[9].

O quarto *corpus* é referido de modo vago por orientandos da Professora Catarina de Sena (UFPI). São citadas realizações fonológicas não-padrão extraídas da fala de seis informantes, moradores da comunidade de Vila da Paz, em Teresina (GOMES et al., 2013).

O quinto *corpus* foi organizado por autor. Intitula-se *Fala Teresinense: Recontato.* Compõe-se de 11 entrevistas de 1 hora, realizadas em 2015. Os informantes são um subgrupo daqueles entrevistados no Projeto *Aspectos Gramaticais do Português Falado por Estudantes Teresinenses.* Foram registrados a idade, nível de escolaridade, profissão e sexo de cada informante. A gravação digital foi realizada usando-se MP3. Pretende-se disponibilizar esse *corpus* na Internet. Por enquanto esse material está sendo preparado de modo a omitir dados que possam vir a identificar o informante e ainda não foi totalmente transcrito. A constituição das entrevistas, a temática e as normas de transcrição são as mesmas do *Corpus Português falado por estudantes teresinenses*, já referido (PESSOA, 2015).

Um quadro síntese sobre os corpora referidos até aqui aparece a seguir. Na primeira coluna encontra-se um conjunto de informações minimamente necessárias para se identificar cada *corpus* e seu estágio atual de construção. Esse conjunto poderia ser ampliado e refinado, sendo, portanto, preliminar. Mas, para os nossos objetivos, é suficiente na medida em que pretende informar sobre o andamento da construção de *corpora* e mostrar o quanto essa tarefa é necessária, ainda que o trabalho esteja em fases iniciais. Os itens da primeira coluna podem ser vistos como etapas de um roteiro, apesar de sua simplicidade. As demais colunas do Quadro são assinaladas com "sim", "parcial" ou "não" em relação a cada item. O ponto de interrogação indica não obtenção de resposta à pergunta formulada.

[8] Soma-se a esse corpus um outro gravado em 2008, composto de nove frases produzidas por 20 pessoas de idades variadas, que vão de 16 a 60 anos, contando, também, com grau de escolaridade variada. Todos eram naturais de Teresina de cidades circunvizinhas à capital, preferencialmente, que se situam a norte do Estado. A gravação foi realizada no Estúdio de Rádio e Produção no Centro de Ensino Unificado de Teresina – CEUT, monitorado pelo técnico Benício Franco, diretamente no computador Pentium IV, com 2HD de 20Gb, 256Mb de memória RAM e processador de 700Mhz. Foram utilizados ainda: mesa de som Stanner, microfone Somson; placa de áudio Delta 44 com ajuda do software Sound Forge 6.0. A partir da placa de som do computador que gerencia o estúdio, o corpus foi gravado diretamente em CD-ROM, de modo que a cada repetição correspondesse uma faixa. O total de frases gravadas é 180. O conjunto de frases que foi objeto de análise acústica por Carvalho (2008) é: (1) Eu falei porta duas vezes;(2) Eu falei corta duas vezes;(3) Eu falei morta duas vezes;(4) Eu falei porte duas vezes;(5) Eu falei corte duas vezes;(6) Eu falei morte duas vezes;(7) Eu falei posta duas vezes;(8) Eu falei costa duas vezes;(9) Eu falei mostra duas vezes.

[9] Em Oliveira Silva e Carvalho (2012) analisa-se um corpus de fala de 12 informantes (06 homens e 06 mulheres), com faixa etária variando entre 25-49 anos e mais de 50 anos, e nível de escolaridade variando entre analfabeto, fundamental e médio. Em Nunes, Barros do Carmo e Carvalho (2011, p.37) há referência a um corpus de fala, gravado, de "[...] teresinenses do sexo masculino e feminino com faixa etária variando entre 30 a 50 anos e mais de 50 anos, com e sem escolaridade". Nunes e Silva (2011, p.67) mencionam um "[...] banco de dados de fala coletados a partir de entrevistas pessoais com 36 falantes teresinenses estratificados socialmente por gênero (masculino e feminino), faixa etária (20-35; 36-50 e + 50 anos) e escolaridade (Ensino Fundamental, Ensino Médio e Ensino Superior)".

Quadro 1: Situação atual dos *corpora* sobre o dialeto de Teresina (PI)

	Corpus 1[10]	*Corpus* 2[11]	*Corpus* 3[12]	*Corpus* 4[13]	*Corpus* 5[14]
Dimensão atual e dimensão pretendida	96/96	36/36	36/36	12/12	11/11
Atribuição de um título ao *corpus*	Sim	Não	Não	Não	Sim
Explicitação da data de coleta de dados	Sim	Sim	Sim	?	Sim
Termo de Consentimento Livre e Esclarecido	Sim	Sim	Sim	?	Sim
Supressão de nomes e outros identificadores potenciais do entrevistado	Sim	?	?	?	Sim
Transcrição dos dados	Sim	Sim	Sim	?	Parcial
Digitalização das transcrições	Sim	Sim	Sim	?	Não
Digitalização do áudio	Sim	Sim	Sim	?	Sim
Indicação de "como citar"	Sim	Não	Não	Não	Sim
Informatização do *corpus* a) Etapas de preparação dos textos b) Disponibilização de ferramentas de manipulação dos dados	Não Não	Não Não	Não Não	Não Não	Não Não
Disponibilização da transcrição	Irrestrita (livro)	Restrita	Restrita	Restrita	Restrita
Disponibilização do áudio dos textos orais	Não	Não	Não	Não	Não
Endereço eletrônico	?	?	?	?	?

Este quadro mostra resultados importantes na medida em que fornece um panorama do estágio atual dos *corpora* de fala teresinense. Apenas 20% do total dos *corpora* tiveram as transcrições publicadas e nenhum disponibilizou livremente os áudios até o momento. Quatro dos cinco *corpora* ainda esperam por título e aguardam ser reconhecidos como um produto relevante dentre outros gerados pelos projetos de pesquisa. É importante ressaltar o atendimento à exigência de assinatura do termo de consentimento por parte dos organizadores dos *corpora* e o cumprimento quase total da etapa de digitalização do áudio e das transcrições, o que viabiliza uma futura disponibilização e publicação digital. Veja-se que a situação exposta neste Quadro não é um caso único e nem tampouco um caso muito diverso de outros identificados e identificáveis no Brasil.

[10] Lima;Serra (2010).

[11] Silva(2009)

[12] Carvalho(2009)

[13] Costa;Viana,Araujo,Sila;Costa(2013)

[14] Pessoa (2015)

4 POR UMA POLÍTICA DE AUTORIA

As tarefas de planejamento, construção e disponibilização de um *corpus* são onerosas ao pesquisador. A visibilidade desse longo trabalho fica obscurecida pelos resultados de análise de processos gramaticais e tratamento estatístico. Por que apenas 20% dos corpora do Quadro I possuem um título, sendo registrados como um trabalho independente nas referências dos artigos em que foram utilizados? Conforme veremos mais adiante, esse baixo percentual pode ser atribuído ao não reconhecimento da formação de um *corpus* como um trabalho autônomo e relevante.

Uma vez formado um *corpus*, é muito importante prepará-lo para que possa ser visto como um produto e possa servir a novas pesquisas. Muitas vezes, os *corpora* permanecem engavetados, o que representa subaproveitamento de recursos humanos e financeiros. Por ser um produto, é necessário que receba um título, tal como qualquer outra produção. Mas não é só isso. Informar o nome do organizador, fornecer uma descrição da dimensão do *corpus* e do perfil dos informantes é fundamental. Apresentar com clareza uma instrução de como citar o *corpus* é outra informação essencial para se garantir a autoria do trabalho. Um caso exemplar encontra-se no site do *Corpus* do Português.

> *Corpus* do Português -- de 20 milhões de palavras (do século XX) a um bilhão (mil milhões) de palavras (do século XXI).
>
> COMO CITAR: DAVIES, M.; et al. (2006) *Corpus do Português: 45 million words, 1300s-1900s*. Available online at http://www.corpusdoportugues.org/.

A importante questão de uma política de autoria é focalizada por Barbosa, Lopes e Callou (2002, p.31) no que diz respeito ao tratamento de textos antigos.

> Não se pode esquecer do tempo investido até que os textos estejam preparados para o levantamento e análise linguísticos. Entre a localização dos documentos em acervos privados até o aparato crítico são consumidas grandes parcelas de tempo de trabalho do linguista [...]. Para que o investimento na etapa filológica não seja desconsiderado, é necessário assumir a autoria individual de cada fonte editada: desde o momento em que se estabelece a conexão com o site, é necessário anunciar que qualquer uso deverá citar 1) o membro da equipe responsável pela edição das fontes usadas; 2) o acervo onde estão os originais; e, por fim, 3) a equipe.

Parafraseando esses autores, também os organizadores de corpora sincrônicos de língua falada não podem esquecer do tempo investido até que as entrevistas estejam preparadas para o levantamento e análise linguísticos. Entre a localização dos informantes, a realização das gravações, as transcrições e preparação do material são consumidas grandes parcelas de tempo de trabalho do linguista. Para que o investimento na etapa de preparação de textos falados não seja desconsiderado, é necessário registrar a contribuição individual de cada pesquisador e auxiliar de pesquisa. É necessário anunciar que qualquer uso deverá citar: 1) o nome do organizador; 2) o título do *corpus*; e 3) o endereço onde localizá-lo. Seria também muito oportuno não omitir os nomes dos entrevistadores e transcritores, sempre que possível. Esse trabalho, geralmente anônimo, é muitas vezes prestado por alunos na fase de iniciação científica. O trabalho de revisão também não pode ser esquecido.

Em resumo, dentre as várias etapas de formação de um *corpus* deve-se incluir a de divulgação do trabalho realizado. Outra etapa também importante é a de disponibilização irrestrita. Para a disponibilização, é importante ressaltar as exigências de exclusão de nome e de informações que poderão permitir a identificação do informante. Outras tarefas, que vão variar conforme o tipo de banco de dados pretendido, são a etiquetagem, os cabeçalhos, a descrição da situação de coleta, perfil do entrevistador, dentre outras[15]. O reconhecimento da complexidade das tarefas envolvidas na organização de um *corpus*, justifica tratá-lo como um trabalho independente e não como uma etapa metodológica de um projeto.

5 OUTROS *CORPORA*

Além dos *corpora* do Quadro I, cuja metodologia é variacionista, é oportuno mencionar dois *corpora* de fala teresinense cujas formações resultam de outras metodologias. Parece-nos que diferenças na metodologia de coleta não são, em si, um intransponível obstáculo para a análise sociolinguística de um conjunto de entrevistas. Talvez muito trabalho fosse poupado se os sociolinguistas adotassem uma prática interdisciplinar, aproximando-se de pesquisadores de áreas afins. Vejamos dois exemplos.

Marques-Ribeiro (2007), em sua dissertação sobre práticas de leitura em língua estrangeira, constrói um *corpus* de 30 entrevistas realizadas em *lanhouses*, cujos sujeitos são pré-adolescentes, alunos de escolas públicas, nível fundamental, usuários de jogos eletrônicos (VG) (Ver transcrição de parte dessas entrevistas à página 171).

Silva (2012, p.22) apresenta um *corpus*, formado por entrevistas, no seguinte formato:

> Questões abertas em que os entrevistados se posicionaram de forma livre, com o emprego do gravador e registros fotográficos e em diários de campo. Foram entrevistados mestres [...] de capoeira em Teresina, além de pessoas que de alguma forma vivenciaram essa história, contribuindo com seus relatos e testemunhos [...]. As entrevistas foram transcritas

As transcrições de duas dessas entrevistas estão disponíveis nas páginas 187-192 e 217-220 desse trabalho. As entrevistas foram realizadas no próprio local de moradia do informante. Foi feita transcrição ortográfica acompanhada de transcrição fonética parcial. Embora esse *corpus* tenha sido feito em conformidade com metodologias de história oral[16], o resultado é de interesse para um sociolinguista.

6 CONCLUSÕES

Neste artigo buscamos fazer um estudo de caso, tomando como objeto um conjunto de iniciativas que visam a registrar a língua portuguesa tal como é falada na cidade de Teresina (PI). Vimos que se trata de um dialeto que recebeu um número relativamente pequeno de estudos sociolinguísticos. Entretanto, a contribuição desses estudos já é suficiente para a descrição de numerosos processos sintáticos, lexicais, morfológicos, discursivos e fonológicos, na medida em que, juntos, conseguiram reunir quase duzentas horas de gravação de fala natural, digitalizadas, de informantes de faixas etárias distintas. Se esses dados pudessem ser reunidos de modo a formar um banco de dados, todos seríamos beneficiados. Esse total de horas representa um conjunto de aproximadamente 1.500.000 palavras.

[15] Sobre etapas metodológicas para a compilação de um corpus linguístico e enumeração de ferramentas computacionais, além de exemplos de bancos de dados sobre o português brasileiro, ver Aluísio e Almeida (2006).

[16] [...] a história oral é um método de pesquisa (histórica, antropológica, sociológica, etc.) que privilegia a realização de entrevistas com pessoas que participaram de, ou testemunharam, acontecimentos, conjunturas, visões de mundo, como forma de se aproximar do objeto de estudo. Como consequência, o método da história oral produz fontes de consulta (as entrevistas) para outros estudos, podendo ser reunidas em um acervo aberto a pesquisadores. Trata-se de estudar acontecimentos históricos, instituições, grupos sociais, categorias profissionais, movimentos, conjunturas etc. à luz de depoimentos de pessoas que deles participaram ou testemunharam." (ALBERTI, 2005, p.18 apud SILVA, 2012, p. 138).

Feito nosso estudo de caso, argumentamos a favor do reconhecimento da organização de um *corpus* como um trabalho independente e não apenas o atendimento a uma etapa metodológica da pesquisa. Buscamos justificar a relevância de fornecer de forma clara instruções de como citar um *corpus* e de divulgar esse tipo de trabalho via internet.

Buscamos oferecer, no Quadro 1, um conjunto de quesitos a serem atendidos, de modo a permitir informar, ainda que minimamente, sobre o andamento da formação de um *corpus*. Nossas sugestões poderiam ser vistas como um primeiro passo para viabilizar a construção de um quadro geral dos *corpora* já formados e daqueles em formação, fornecendo assim uma fotografia atualizada, em que se indicaria endereço, forma de acesso, etc. Esse quadro geral poderia ser disponibilizado num site já conhecido dos sociolinguistas, como o do GT de Sociolinguística da ANPOLL.

REFERÊNCIAS

ALBERTI, V. *Manual de história oral*. 3.ed. Rio de Janeiro: Editora Fundação Getúlio Vargas. 2005.

ALUÍSIO, S. M.; ALMEIDA, G. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calidiscópio*, v., n.3, p.156-178, 2006.

BARBOSA, A.; LOPES, M.R.; CALLOU, D.. Organização dos corpora diacrônicos do PHPB-RJ na rede mundial de computadores. In: DUARTE, M. E. L.; CALLOU, D. (Org.). *Para a história do português brasileiro*. Rio de Janeiro: Ed. FAPERJ, 2002, vol. 3. p. 29-38.

CARNEIRO, J. R. D. *Povos e línguas indígenas no Maranhão*: contato linguístico. 2014. 260f. Tese (Doutorado) – Programa de Pós-Graduação em Linguística, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

CARVALHO, L. S. *Os róticos em posição de coda:* uma análise variacionista e acústica no falar piauiense. 2008. 267f. Tese (Doutorado) – Programa de Pós-Graduação em Linguística, Universidade Federal da Paraíba, João Pessoa, 2008.

CASTILHO, A. O português culto falado no Brasil. História do Projeto NURC/BR. In: PRETI, D.; URBANO, H.. (Org.). *A linguagem falada culta na cidade de São Paulo*. São Paulo: T. A. Queiroz, Fapesp, 1990, vol. IV. p. 141-202.

DANTAS, J.; CARVALHO, M.; COSTA, C. Variação linguística versus ensino de língua portuguesa em nossas escolas: como erradicar o círculo vicioso do preconceito linguístico? *Sociodialeto*, v.5, n. 14, p. 290-305, 2014.

DAVIES, M. et al. *orpus do Português: 45 millionwords, 1300s-1900s*. 2006. Disponível em: <http://www.corpusdoportugues.org>. Acesso em: 07 jul. 2016.

FREITAG, R. M. K. *Metodologia de coleta e manipulação de dados em sociolinguística*. São Paulo: Blucher, 2014.

GOMES, B. R. I. S. et al. A influência do padre Pedro Balzi na fala dos moradores da comunidade de Vila da Paz. *Sociodialeto*, v. 4, n. 11, p. 1-10, 2013.

GUMPERZ, J. HYMES, D.. *Directions in sociolinguistics*. New York: Holt, Rinehart and Winston, 1972.

IBGE. Fundação Instituto Brasileiro de Geografia e Estatística. Censo demográfico 2010. Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2010/default.shtm>. Acesso em: 05 mar. 2015.

LABOV, W. *Padrões sociolinguísticos.* Trad. Marcos Bagno, Maria Marta Pereira Scherre, Caroline Rodrigues Cardoso. São Paulo: Parábola Editorial, 2008 [1972].

LIMA, M. A. F.; SERRA, M. A. C. M. *Português falado por estudantes teresinenses*. Teresina: EDUFPI, 2010. vol. I.

LIMA, P. H. G. *A ocorrência de Policentralidade em Teresina* – PI: a formação de um subcentro urbano na Região Sudeste. 2011. 204f. Tese (Doutorado) – Instituto de Geociências e Ciências Exatas, Universidade Paulista Júlio Mesquita, Rio Claro, 2011.

MARQUES-RIBEIRO, A. J. P. *Práticas de leitura do gênero de discurso videogame*. 2007. 172f. Dissertação– Programa de Pós-graduação em Letras, Universidade Federal do Piauí. Teresina, 2007.

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

NUNES, A.; BARROS do CARMO, W.; CARVALHO, L. *A ocorrência de metaplasmos em textos da língua falada*: *uma análise sociolinguística.* In: SIMPÓSIO DE PRODUÇÃO CIENTÍFICA E XI SEMINÁRIO DE INICIAÇÃO CIENTÍFICA, 7., Teresina, 2011. Livro de Resumos... Teresina: UESPI, 2011. Disponível em: <http://www.uespi.br/prop/XII_SIMPOSIO_2012/Livro_de_Resumos_XII%20SIMPOSIO%20PRONTO2.pdf>. Acesso em: 20 set. 2016.

NUNES, F. G. A. SILVA, A. A palatalização do fonema /s/, em posição de coda no falar teresinense. In: SIMPÓSIO DE PRODUÇÃO CIENTÍFICA, 11.; SEMINÁRIO DE INICIAÇÃO CIENTÍFICA, 11., Teresina, 2011. *Livro de Resumos...* Teresina: Universidade Estadual do Piauí, 2011. p.67

OLIVEIRA SILVA, L. R. CARVALHO, L.S.*Um estudo sobre a despalatalização no falar teresinense: uma análise sociolinguística*. In: SIMPÓSIO DE PRODUÇÃO CIENTÍFICA, 7.; SEMINÁRIO DE INICIAÇÃO CIENTÍFICA, 11., Teresina, 2012. *Livro de Resumos...* Teresina: Universidade Estadual do Piauí., 2012. p.167.

PESSOA, F. C. R. Estudo em painel do uso do dativo na fala de Teresina. In: SEMANA DE EVENTOS DA FACULDADE DE LETRAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS (SEVFALE). Belo Horizonte, 2015. *Livro de Resumos...* Belo Horizonte: UFMG, 2015. p.119-127. Disponível em: <http://anais.letras.ufmg.br/index.php/SEVFALE/XIISEVFALE/paper/view/17/13>. Acesso em: 07 jul 2016 .

ROMAINE, S. A critical overview of the methodology of British urban sociolinguistics. *English World Wide* , v. 1, Oxford , p.163-99, 1980.

SILVA, A. N. *As pretônicas no falar teresinense*. 2009. 236f. Tese (Doutorado) – Programa de Pós-Graduação em Letras, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2009.

SILVA, R. C. *As narrativas dos mestres e a história da capoeira em Teresina/PI:* do pé do berimbau aos espaços escolares. 2012. 308f. Tese (Doutorado) –Programa de Pós-Graduação em Educação Brasileira, Universidade Federal do Ceará, Fortaleza, 2012.

SOUSA, V. R. Discutindo a relação sociolinguística: uma análise semântica da linguagem de Angical do Piauí (PI). In: SIMPÓSIO INTERNACIONAL DE ENSINO DE LÍNGUA PORTUGUESA. Uberlândia, 2012. **Caderno de Resumos**..., Uberlândia 2012.

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

Recebido em 06/09/2016. Aceito em 25/10/2016.