CAPÍTULO 1

DESAFIOS DA CONSTITUIÇÃO DE CORPORA LINGUÍSTICOS

Silvia Figueiredo Brandão
Universidade Federal do Rio de Janeiro/CNPq

1 INTRODUÇÃO

Embora já haja significativo número de bancos de dados linguísticos, que têm concorrido para a caracterização das modalidades oral e escrita do Português do Brasil (PB) dos pontos de vista sincrônico e diacrônico, muitos são, ainda, os desafios quando se trata de constituir ou ampliar corpora, como têm constatado todos os que já passaram por essa experiência e como vêm demonstrando pesquisadores que se vinculam à área da Linguística de Corpus.

Deixando à parte, entre outros, os dados recolhidos em obras precursoras, como *O dialeto caipira*, de Amadeu Amaral (1920) e *O falar carioca*, de Antenor Nascentes (1922), destaca-se o projeto do Atlas Prévio dos Falares Baianos (APFB) (ROSSI et al., 1963). Sua realização remonta a meados da década de 1950 – que iniciou, no Brasil, a prática de pesquisa de campo e de organização de corpora com base em coleta rigorosa e metodologia bem definida, numa perspectiva essencialmente diatópica e sem os recursos hoje disponíveis para registro dos dados. Mais tarde, na década de 1970, tem início a formação de corpora de perfil sociolinguístico, embora com motivações e objetivos bem diferentes.

O corpus organizado pelos pesquisadores vinculados ao NURC, surgiu no âmbito de um projeto internacional que tinha por meta observar a fala de indivíduos de nível superior de instrução no intuito de retratar a modalidade oral culta, no caso brasileiro, nas cinco cidades que, à época, tinham mais de um milhão de habitantes: Rio de Janeiro, São Paulo, Recife e Salvador, fundadas no século XVI, e Porto Alegre, no século XVIII. O projeto nasceu, em 1968, da sugestão de Nelson Rossi de estender ao Brasil o *Proyecto de Estudio Coordinado de la Norma Lingüística Culta de las Principales Ciudades de Iberoamérica y de la Península Ibérica*, por serem "tão evidentes e tão relevantes os pontos comuns à problemática do espanhol nas Américas e do português no Brasil" (ROSSI apud https://nurcrj.letras.ufrj.br/).

Para a constituição do corpus, que se iniciou em 1972, foram estabelecidos três tipos de gravações: aulas e conferências (Elocuções formais/EF), diálogos informais (Diálogos entre dois locutores/D2) e entrevistas (Diálogos entre locutor e documentador/DID), com indivíduos naturais das cidades-alvo, de ambos os sexos, distribuídos por três faixas etárias (de 25 a 35 anos, de 36 a 55 e 56 anos em diante). Com a implementação do Projeto Gramática do Português falado, em finais da década de 1990, 18 entrevistas de cada cidade (três por célula social) serviriam de base para as análises da equipe que o constituía.

O projeto Competências básicas do Português, coordenado por Miriam Lemle e Anthony Naro foi o ponto de partida do que mais tarde seria denominado de Programa de Estudos sobre o Uso da Língua (PEUL), cujos integrantes muito colaboraram com orientações para a formação de outros corpora, como os do VARSUL e do VAL-PB.

O projeto inseria-se no âmbito do Movimento Brasileiro de Alfabetização (MOBRAL) e tinha por objetivo principal "verificar pontos de diferenciação entre a variedade de língua portuguesa falada por esse grupo social [os alfabetizandos] e as variedades de língua escrita nas quais almejam eles adquirir competência", de modo a contribuir para uma "gradação adequada do material didático" a eles dirigido (LEMLE; NARO, 1977, p. 2).

Para viabilizar a pesquisa, que contou também com uma amostra de língua escrita (de histórias em quadrinhos, fotonovelas, jornais classes A e B e literatura nacional) foram realizadas entrevistas com 20 alunos do MOBRAL, naturais do Grande Rio, 9 mulheres e 11 homens, 6 deles com mais de 40 anos e 14 com menos de 30. Como havia a intenção de testar o fator estilístico de grau de formalidade, as entrevistas com cada informante eram realizadas em várias etapas: no local das aulas (as 1ª, 2ª e 5ª), na residência de diferentes entrevistadores (as

3ª e 4ª), no local das aulas ou na residência do entrevistado, com um trecho de gravação à revelia (a 6ª), no local das aulas ou na residência do entrevistado, em companhia de algum de seus amigos (a 7ª). Os informantes tinham liberdade para falarem sobre temas de sua preferência, embora lhes tenha sido apresentada uma lista de tópicos de "possível interesse antropológico", sugerida por Roberto da Matta (LEMLE; NARO, 1977, p. 5-7).

Ao corpus Censo, como foi denominada a amostra inicial, incorporaram-se outras do português falado, com informantes de ambos os sexos distribuídos por faixa etária e pelos níveis fundamental e médio de escolaridade, e, em virtude de novas questões teóricas que se iam impondo, amostras de língua escrita antiga e contemporânea, e, ainda, de Português de contato.

Com o passar do tempo, tais iniciativas, impulsionadas pela rápida difusão da Sociolinguística de inspiração laboviana, desencadeariam em diversas áreas do país, o interesse pela organização de bancos de dados e permitiriam testar hipóteses sobre a difusão e as possíveis implicações sociais de fenômenos linguísticos variáveis.

Para tratar dos desafios na constituição de corpora, sem esquecer exemplos do muito que se conquistou com base naqueles de que já se dispõe, parte-se de algumas questões presentes em Sardinha (2000), em experiências pessoais e em depoimentos e textos de colegas que já organizaram corpora. Acredita-se que, apesar dos avanços na área dos estudos linguísticos e dos recursos da informática, algumas das dificuldades inerentes à composição e divulgação de bancos de dados sejam uma constante preocupação de sociolinguistas e de dialetólogos, bem como de outros pesquisadores que desenvolvem estudos de base empírica. Tais dificuldades, no que respeita a bancos de dados de perfil sociolinguístico, decorrem, por vezes, não só de questões teórico-metodológicas advindas de novas tendências quer da Dialetologia, quer da Sociolinguística, como aponta Freitag (2015), mas também derivam de questões de ordem prática, tais como, a escolha de ferramentas para disponibilização e compartilhamento de corpora.

Nesse sentido, este texto tem continuidade em duas outras seções – a primeira, dedicada à noção de corpus (item 1), a segunda, às conquistas e desafios relacionados à constituição de bancos de dados (item 2) – a que se somam as considerações finais (item 3).

2 A NOÇÃO DE CORPUS

Aluísio e Almeida (2006) procuram estabelecer uma diferença entre a concepção de corpus para linguistas em geral e para pesquisadores que se vinculam à área da Linguística de Corpus. Para caracterizar a primeira delas, citam, entre outros, Ducrot e Todorov, para os quais, corpus seria "um conjunto, tão variado quanto possível, de enunciados efetivamente emitidos por usuários da referida língua, em determinada época" (DUCROT; TODOROV, 2001 apud ALUÍSIO; ALMEIDA, p. 157). Já para aqueles que se vinculam à área da Linguística de Corpus, as amostras devem estar sempre em formato eletrônico.

Como, atualmente, a tendência é viabilizar o uso de ferramentas computacionais para tornar mais fácil a busca de informações em bancos de dados, parte-se da definição de Sanchez, um linguista de corpus (1995 apud SARDINHA, 2000, p. 8-9), que os caracteriza como

> um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para descrição e análise (SANCHEZ, 1995 apud SARDINHA, 2000, p. 8-9).

Nesse conceito de corpus, representatividade parece ser a palavra-chave, por estar, segundo Sardinha (2000, p. 342-348), intimamente ligada à sua extensão, quer em virtude de a linguagem ser um sistema probabilístico (pois há traços mais e menos frequentes), quer pelo fato de um corpus ser uma amostra de um todo (uma população linguística, como ele denomina), cuja real dimensão não se conhece. Desse modo, a sua representatividade seria sempre relativa, o que dependeria não só da resposta à pergunta representativa de quê, mas ainda representativa para quem, isto é, os usuários de um corpus teriam o ônus de "demonstrar a representatividade da amostra e de serem cuidadosos em relação à generalização dos seus achados no que toca ao todo (um gênero ou a língua inteira, por exemplo)". A representatividade do corpus determinaria, segundo Biber, Conrad e Reppen (1998), os tipos de questões que podem ser formuladas e a generalização dos resultados advindos da pesquisa.

3 CONQUISTAS E DESAFIOS

Passa-se, agora, a tratar, com ênfase na perspectiva sociolinguística, inicialmente, de desafios de ordem teórico-metodológica, com base na noção de

representatividade e comparabilidade (2.1), para, em seguida, focalizar desafios que dizem respeito a disponibilização, centralização e compartilhamento, nem sempre consensual, de corpora linguísticos (2.2), com base nas primeiras iniciativas de debate sobre esses temas no âmbito da ABRALIN, bem como de questões propostas por Freitag em recente *live* por ela mediada.

3.1 Questões de ordem teórico-metodológica

3.1.1 No âmbito de comunidades de fala

Não resta dúvida de que, nos últimos 70 anos, tomando como ponto de partida o APFB, a organização de corpora redundou em inúmeras pesquisas sobre diferentes comunidades de fala. Além dos atlas já publicados e de outros desenvolvidos, como teses e dissertações, sobretudo a partir de 1996, quando se instalou o comitê encarregado de elaborar o ALiB, projetos, como o NURC, o PEUL, o VARSUL, o ALIP, o VALPB, o FALA-NATAL, o SP2010, o APERJ, o COMPARAPORT, entre vários outros, vêm fornecendo valiosas descrições de variedades cultas e populares do PB.

Devem-se destacar, ainda, corpora representativos de variedades não brasileiras: o Corpus Concordância, reúne, além de entrevistas relativas ao PB (realizadas em Copacabana e em Nova Iguaçu, na região Metropolitana do Rio de Janeiro) as gravadas em duas localidades da região Metropolitana de Lisboa (Oeiras e Cacém) bem como, as realizadas por Tjerk Hagemeijer em São Tomé e Príncipe. O Corpus Moçambique-Port, recolhido por Vieira e Pissurno (2016) em Maputo conta com 35 entrevistas a que se juntarão as que estão sendo realizadas remotamente.[1] Por sua vez, o corpus do Projeto Em busca das raízes do português brasileiro compõe-se de entrevistas realizadas em Luanda.

Como observa Brandão (2013, p. 3):

> Os conhecimentos já existentes sobre a estrutura e a dinâmica das línguas em diversas perspectivas teóricas bem como as descrições sócio e geolinguísticas de que já se dispõe se, de um lado, podem servir de parâmetros para a realização de novas amostras, de outro, requerem que o pesquisador não só tenha clareza quanto à possibilidade de encontrar bem representadas as estruturas/variáveis que objetiva analisar, mas também que considere as características demográficas, socioeconômicas e culturais da área focalizada. Cabe, portanto, decidir, no caso de um corpus oral, sobre a metodologia de coleta, o número de informantes e os parâmetros que presidirão à sua seleção (BRANDÃO, 2013, p. 3).

[1] As amostras relativas ao Português do Brasil, ao Português Europeu e ao Português de Moçambique podem ser acessadas em https://corporaport.letras.ufrj.br.

A grande dificuldade em retratar adequadamente a realidade linguística brasileira advém da complexidade e da heterogeneidade sociocultural do país, devida, de um lado, ao maciço contato multilinguístico e multiétnico que se verifica desde a época da colonização, de outro, a um processo de deslocamento de grande parte da população rural para áreas urbanas.

No Censo Geral do Império, de 1872, o primeiro realizado no país, o Brasil contava com 9.930.478 habitantes, dos quais apenas 10,41% nas capitais provinciais e no Município Neutro,[2] sendo que, deste percentual, 48% se concentravam nas cidades do Rio de Janeiro, Salvador e Recife.[3] A transformação do "vasto país rural" nas palavras de Celso Cunha, em vasto país urbano fica mais evidente a partir dos anos 1940 do século XX.

A mudança de perfil foi abrupta em algumas regiões, como no Centro-Oeste (21,5% para 88,8%) e gradativa no Norte e Nordeste. No Sudeste, onde sempre se encontraram as taxas mais altas de urbanização, o aumento, em 70 anos, foi de cerca de 53%, devendo-se, ainda, levar em conta que tais deslocamentos não se deram apenas no âmbito intrarregional, tendo havido, ainda,

> [...] deslocamentos inter-regionais, determinados pela busca de melhores oportunidades de trabalho. Na Região Sudeste, em que se localizam as duas maiores regiões metropolitanas do país, a de São Paulo, com 20.935.204 habitantes, e a do Rio de Janeiro, com 11.973.505, é onde a concentração de população urbana atinge o maior índice: 99,3%. Grande parte desse contingente é oriundo de outras regiões do país, sobretudo do Nordeste e de Minas Gerais, o que torna esses espaços extremamente complexos não só do ponto de vista social, mas também linguístico, uma vez que neles se configura um intenso contato interdialetal (BRANDÃO, 2015, p. 201).

Nas pesquisas tradicionais que visam à caracterização de uma determinada área, os informantes são naturais da localidade pesquisada bem como seus pais, o que, em zonas urbanas, por vezes, torna difícil a seleção de informantes. Na amostra do Atlas Fonético do Entorno da Baía de Guanabara (AFEBG) (LIMA, 2006), que contempla quatro comunidades da região Metropolitana do Rio de Janeiro, foi necessário flexibilizar esse critério tendo em vista que a grande maioria da população das localidades-alvo era natural de outras cidades fluminenses ou de outros pontos do país, sobretudo do Nordeste, de Minas Gerais e do Espírito Santo. Assim, convencionou-se que o informante poderia ser oriundo de outra

[2] O Município Neutro, unidade administrativa criada em 1834, corresponde ao atual Município do Rio de Janeiro, que passou a ser denominado de Distrito Federal de 1891 até 1960, quando Brasília se tornou a capital do Brasil e, consequentemente, Distrito Federal.

[3] https://pt.wikipedia.org/wiki/Censo_demogr%C3%A1fico_do_Brasil_de_1872. Acesso em: 3 nov. 2020.

localidade, desde que tivesse ido morar na comunidade-alvo com até cinco anos de idade, desconsiderando-se a exigência inicial quanto à naturalidade dos pais. Dos 24 informantes, 7 procediam de outras localidades. Quanto à mães e pais dos informantes, respectivamente, 15 (62,5%) e 17 (70,8%) não tinham nascido nas localidades-alvo.

No site do Projeto SP2010, embora os critérios gerais para seleção de informantes tenham levado em conta sexo/gênero, faixa etária, escolaridade, região e zona da cidade, anotaram-se dados sobre a naturalidade dos pais e avós dos entrevistados.

Bortonni-Ricardo (2004) indica como um dos eixos fundamentais para o estudo da variação no PB, o que ela denominou de *continuum* de urbanização que tem como polos variedades rurais isoladas e variedades urbanas padronizadas intermediadas pela zona rurbana.

Entre os corpora que contemplaram comunidades rurais encontram-se os que compõem o projeto Vertentes do Português Popular da Bahia[4] que conta com entrevistas em quatro comunidades afrodescendentes, dois municípios do interior, e, ainda, amostras de fala de quatro bairros populares de Salvador e de um município de sua região Metropolitana. Outro projeto, A Língua Portuguesa no Semiárido Baiano, nas suas duas primeiras fases, voltou-se para a constituição de corpora representativos de comunidades rurais (1996-2001) e, a partir de 2007, a Feira de Santana, havendo, ainda, a intenção de estudar o português falado em áreas indígenas especiais.[5]

Como observa Souza (2009, p. 181), entre os espaços rurais e urbanos, "as fronteiras esmaecem, seus contornos, outrora nítidos, borram-se, tornam-se imprecisos; dilatam-se e esfacelam-se em inúmeras situações intermediárias", constituindo o que se convencionou denominar de espaços rurbanos, por sua vez, também difíceis de caracterizar.

Espaços rurbanos poderiam ser a periferia das cidades e as favelas, por vezes incrustadas nas chamadas áreas nobres de grandes centros urbanos, como a Rocinha, em São Conrado, Cantagalo-Pavão-Pavãozinho, na fronteira entre Ipanema e Copacabana, no Rio de Janeiro, ou Paraisópolis, na região do Morumbi na capital paulista. Nessas comunidades, que, em geral, abrigam indivíduos vindos de outras áreas do país, há, em maior ou menor grau, a manutenção de seus dialetos de origem e a acomodação ao dialeto local a depender de pertencerem

[4] Cf. http://www.vertentes.ufba.br/.

[5] Cf. http://www2.uefs.br/nelp/fases_subprojetos.htm.

a redes sociais mais ou menos densas. Além disso, em algumas comunidades há um nítido sentimento de pertença a um grupo específico, o que leva à adoção de determinadas estruturas linguísticas.

Entre os trabalhos nessa linha, há o de Mollica et al. (2008), que discute, tendo como referência a Favela da Maré, no Rio de Janeiro, "aspectos teórico-metodológicos [...] para analisar os processos migratórios no Brasil [...] e suas repercussões, [...] os modelos através dos quais o contato linguístico é analisado mais adequadamente, em se tratando de comunidades rurbanas localizadas nas periferias das grandes cidades brasileiras" (p. 64). Há, ainda, o trabalho de Rodrigues (2004, p. 120), que tratou da concordância verbal em sua tese com base em 40 entrevistas realizadas em favelas da zona Oeste de São Paulo. Ela comenta:

> [...] instalam-se as dicotomias rural/urbano e culto/popular quando se considera a realidade social da capital paulistana. Nela se verifica um fenômeno especial de variação sociolinguística resultante desse fenômeno de migração interna: a variedade linguística que utilizam os migrantes em seus estados de origem, deixa de representar, simbolizar sua região; tal variedade, regional na origem, torna-se variedade social, símbolo de uma posição social inferior. Os migrantes vão constituir, com a população dessas cidades e de regiões próximas a elas, pertencentes ao mesmo estrato populacional, um extenso grupo de usuários de uma variedade popular ou não-padrão, estigmatizada, que se torna ela mesma um indicador da classe socioeconômica a que pertencem. É lícito esperar que as novas gerações formadas por filhos de migrantes tendam a abandonar os hábitos linguísticos de seus pais, adotando uma variedade de língua que vai, então, refletir a estratificação social urbana e as atitudes sociais que servem para sustentá-la (RODRIGUES, 2004, p. 120).

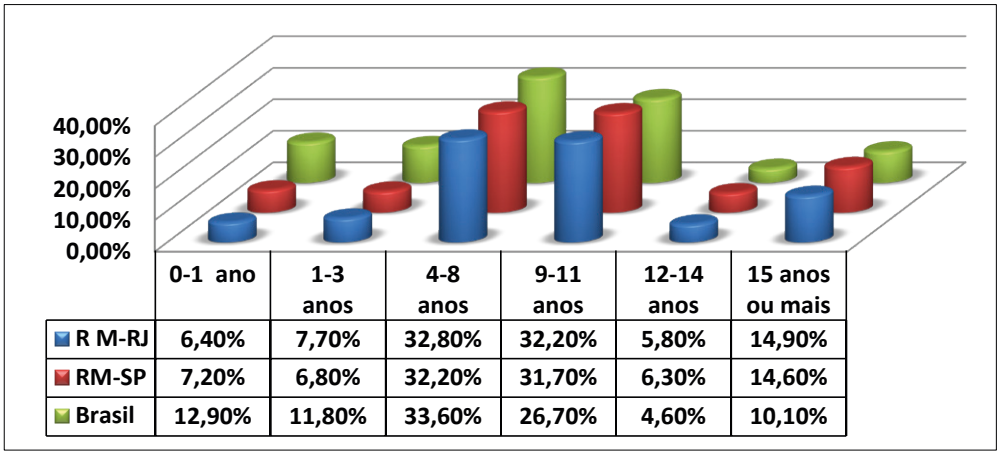
Estão por conhecer também muitas comunidades de médio porte, interioranas. No site do Projeto ALIP (Amostra Linguística do Interior Paulista), afirma-se que uma das motivações para a consecução do Corpus IBORUNA foi a necessidade de "representar o dialeto falado no interior paulista, em razão de este ser ainda pouco conhecido, em bases científicas, por seus usuários, e pelos próprios linguistas".

O *continuum* de urbanização, por sua vez, está associado a nível de escolaridade, parâmetro muito utilizado em estudos sociolinguísticos, em que podem estar embutidas outras variáveis, como o nível socioeconômico, o maior ou menor acesso a bens culturais, maior ou menor contato com indivíduos de diferentes origens geográficas e estratos sociais. Em alguns casos, parece haver sobreposição desse contínuo com o que Brandão (2013a, p. 77) denominou de *continuum* de nível de escolaridade. Em áreas rurais e rurbanas, predominam, via de regra, indivíduos analfabetos ou com baixo índice de escolaridade (0 a 4

ou 5 a 8 anos), enquanto nas grandes cidades há mais probabilidade de se encontrarem indivíduos de nível médio e superior.

Na Figura 1, expõem-se índices percentuais correspondentes a anos de escolarização no Brasil e nas regiões Metropolitanas do Rio de Janeiro (RM-RJ) e de São Paulo (RM-SP).

Figura 1 – Índices percentuais correspondentes a anos de escolarização no Brasil e nas regiões Metropolitanas do Rio de Janeiro (RM-RJ) e de São Paulo (RM-SP).



	0-1 ano	1-3 anos	4-8 anos	9-11 anos	12-14 anos	15 anos ou mais
R M-RJ	6,40%	7,70%	32,80%	32,20%	5,80%	14,90%
RM-SP	7,20%	6,80%	32,20%	31,70%	6,30%	14,60%
Brasil	12,90%	11,80%	33,60%	26,70%	4,60%	10,10%

Fonte: Brandão (2015, p. 203), com base em dados fornecidos pelo IBGE.

Comparando-se as duas regiões metropolitanas, não se observam discrepâncias significativas: os índices percentuais são bastante semelhantes, em todos os níveis. As diferenças se mostram mais claramente quando se comparam os percentuais referentes aos que frequentaram a escola por até 8 anos nas regiões metropolitanas aos do Brasil como um todo. Observe-se, ainda, que os índices correspondentes aos extremos da cadeia de escolarização no Brasil são semelhantes (12,90% que congregam analfabetos e indivíduos que foram à escola por até um ano e 10,10% de indivíduos com nível superior), enquanto nas regiões metropolitanas os percentuais referentes a 15 ou mais anos de escolaridade correspondem ao dobro dos de menor nível de instrução.

Nos grandes bancos de dados, os padrões de distribuição etária dos informantes são variados. Sem mencionar o NURC, cuja repartição por faixas é compatível com o segmento social de falantes com nível superior, no VARSUL, inicialmente, na amostragem urbana, consideraram-se, além de sexo e de três níveis de escolaridade (fundamental I - de 1 a 4 anos, fundamental II - de 5 a 8 anos e nível médio - de 9 a 11), duas faixas etárias: de 25 até 50 e acima de 50 anos. No ALiB, também foram consideradas apenas duas faixas etárias (18 a

30 anos e 50 a 65 anos), opção metodológica que se deveu ao grande número de pontos de inquérito (250) e de informantes, cujo número passaria dos 1.100 considerados para cerca de 1.600, caso se definisse uma faixa etária intermediária.

No projeto FalaPOA, cujo corpus foi organizado entre 2015 e 2019 sob a coordenação de Elisa Batistti, levaram-se em conta indivíduos distribuídos por sexo, três faixas etárias (20-39 anos, 40-59 anos, 60 ou mais anos), três níveis de escolaridade – fundamental, médio, superior –, dois bairros (por renda média mensal em salários mínimos) em cada uma das 4 zonas da cidade (Centro, Norte, Sul, Leste), com base em indicadores sociais e econômicos do site OBSERVA POA (Observatório da Cidade de Porto Alegre).[6]

Sexo/gênero, nível de escolaridade e faixa etária são os parâmetros mais usuais na estratificação dos informantes, embora se venha buscando retratar outras dimensões da variação, diante da já aludida complexidade que caracteriza qualquer comunidade de fala. Dentre elas, estão a variação diafásica ou de registro, contemplada em alguns corpora, como no NURC, com seus três já citados tipos de entrevistas ou no IBORUNA, que, além da Amostra Comunidade (Censo) tem uma amostra de Interação Dialógica, constituída de gravações secretas, na tentativa de fugir ao paradoxo do observador (LABOV, 1972) e chegar ao nível ideal de espontaneidade.

Em alguns bancos de dados de grande porte, por exemplo, as diferentes formas de tratamento são pouco usuais, em função sobretudo do perfil das entrevistas, que, em geral, retratam a interação entre um documentador e um entrevistado. A depreensão das restrições que presidem ao uso de uma ou outra forma, depende não só de diferentes situações intercomunicativas, mas também das relações simétricas/assimétricas que se estabelecem entre os envolvidos na interação, determinadas quer por diferenças relativas à faixa etária, escolaridade e gênero, quer pelos distintos papéis sociais por eles desempenhados. Nesse sentido, seria mais adequado contar com gravações secretas, o que é sempre mais trabalhoso e implica ajustes nos critérios de definição do perfil dos informantes. Para contornar essa lacuna, Célia Regina Lopes, por exemplo, que trabalha com o sistema pronominal, tem recorrido não só a gravações secretas na rua, mas também a filmes.

Novos bancos de dados (e mesmo os antigos, ampliando seus objetivos) deveriam contar com gravações secretas e buscar formas alternativas de gravação de entrevistas por WhatsApp, ou por outros aplicativos, como o Meet ou o Zoom, como têm feito orientandos de Silvia Rodrigues Vieira no sentido de alargar o

[6] Cf. www.observapoa.com.br.

corpus Moçambique-Port. Outro exemplo, é o de Labov, que, em suas recolhas para o *Atlas of North American English* (LABOV; ASH; BOBERG, 2006), contatou os informantes por telefone.

Também a Geolinguística, na perspectiva pluridimensional (RADKE; THUN, 1996) tem buscado integrar à diatopia outras dimensões, entre as quais a diageracional, a diastrática, a diazonal (rural/urbana), a diafásica, a diarreferencial, como indica a proposta de Margotti (2008, p. 2),[7] que, no texto, discute as dificuldades inerentes à apresentação dos resultados que advêm do emprego de métodos da sociolinguística.

Quanto à variação diamésica (variação fala/escrita), o PEUL e o Grupo de Estudos Discurso e Gramática, ambos sediados na UFRJ, dispõem de amostras de fala e escrita. Este último, também atuante na UFF e na UFRN, tem um corpus recolhido em cinco cidades e que obedeceu aos seguintes critérios: cada informante produziu cinco tipos distintos de textos orais e, a partir deles, cinco textos escritos (narrativa de experiência pessoal, narrativa recontada, descrição de local, relato de procedimento, relato de opinião), no intuito de garantir a comparabilidade entre as modalidades falada e escrita. Os informantes, todos alunos, foram distribuídos por níveis de escolaridade/faixa etária, abrangendo desde a alfabetização (de 5 a 8 anos) até o último ano do Ensino Superior – acima de 23 anos.

Outro aspecto relacionado à representatividade retoma a discussão sobre como determinar o número de informantes no corpus como um todo e quantos deles por célula social em análises sociolinguísticas sobre comunidades de fala. Como se sabe, convencionou-se que, em análises variacionistas, o mínimo requerido seriam cinco informantes por célula, o que, se levadas em conta apenas as variáveis clássicas – sexo, faixa etária (3), nível de escolaridade (3) – as análises poderiam requerer, no mínimo, 90 informantes. No entanto, por uma série de motivos, elas, em geral, baseiam-se em dois informantes por célula (por vezes apenas um), o que, apesar disso, tem apresentado resultados consistentes. Os trabalhos realizados com o Corpus Compartilhado do NURC, por exemplo, levam em conta 18 informantes por cidade: 3 homens e 3 mulheres em cada uma das 3 faixas etárias.

Freitag, Martins e Tavares (2012, p. 928), com base no Censo 2010 do IBGE e no índice de 0,5% de representatividade da população estabelecido por Labov em seu estudo sobre Martha's Vineyard, afirmam haver inconsistências nos

[7] Disponível em: http://www.leffa.pro.br/tela4/Textos/Textos/Anais/CELSUL_VIII/geolinguistica_pluridimensional.pdf.

recortes em geral estabelecidos, uma vez que a distribuição por faixas etárias é diferente de área para área e cidades densamente povoadas são representadas pelo mesmo número de informantes que cidades de pequeno e médio portes. Eles exemplificam, no primeiro caso, com dados de Sergipe, em que o maior contingente da população está na faixa de até 24 anos (46,3%) em contraste com a de 40 a 64 (22,9%) e, no segundo, com o Banco IBORUNA, que estipulou o número de informantes segundo a densidade populacional de cada cidade (GONÇALVES, 2008 apud FREITAG; MARTINS; TAVARES, 2012, p. 930).

Uma comparação, por exemplo, com base no Censo de 2010, entre os estados de Sergipe, com 2.068.017 habitantes, e do Rio de Janeiro, com 15.989.929, demonstra que a situação é ligeiramente diferente nas faixas extremas, mas não no que respeita à faixa entre 25 e 39 anos, em que os percentuais são idênticos (24,5%). Sem dúvida, em cada município desses Estados, a situação pode ser diferente, no que é fundamental observar os indicadores sociais ao se definirem critérios de seleção de informantes.

3.1.2 No âmbito de redes sociais e comunidades de práticas

Tendências recentes da Sociolinguística podem também determinar nova dinâmica na constituição de corpora. Freitag, Martins e Tavares (2012) focalizam bancos de dados sociolinguísticos do Português do Brasil, avaliando as suas potencialidades e limitações frente aos estudos da chamada terceira onda. Num quadro (a seguir), eles apresentam, resumidamente, as diferenças entre estudos voltados para a caracterização de comunidades de fala e de comunidades de práticas.

Quadro 1 – Comparação entre abordagens sociolinguísticas de comunidades de fala e de comunidades de práticas.

Abordagem de comunidade de fala	Abordagem de comunidade de práticas
- estratificação baseada em fatores sociodemográficos amplos	- estratificação baseada em valores localmente estabelecidos
- distribuição homogênea, tanto quanto ao tamanho quanto às categorias controladas	- distribuição variável, definida caso a caso
- categorias definidas a priori	- categorias definidas a posteriori
- permissão para captar tendências amplas da comunidade	- permissão para captar valores sociais localmente estabelecidos nas relações
- coleta padronizada (entrevista sociolinguística)	- coleta etnográfica (observação participante, interações entre grupos)
- constituição da amostra em curto prazo	- constituição da amostra em longo prazo

Fonte: Freitag; Martins; Tavares (2012, p. 931).

Como se deduz do Quadro 1, bancos de dados de grande porte, organizados até a década de 2000 (PEUL, VARSUL, NURC, VALPB, entre outros) não servem de base para a análise de comunidades de práticas e de redes de sociais, embora os resultados das análises que Eckert (2012) inclui na primeira onda da Sociolinguística continuem a ser fundamentais para a realização de estudos que se enquadram na segunda e na terceira. Tais estudos, que também operam com dados estatísticos, requerem um trabalho de natureza mais propriamente etnográfica e o convívio do pesquisador com a comunidade, de modo a, detectando como se dá a dinâmica sociolinguística entre os membros de um determinado grupo, determinar como se instaura o significado social da variação.

Há significativas diferenças no que respeita ao perfil do corpus, à metodologia empregada, aos procedimentos analíticos. Para contemplar a nova abordagem em corpora de amplo espectro, os autores sugerem que os bancos de dados incorporem aspectos metodológicos de terceira onda, tais como os que foram implementados no Banco de Fala Culta de Itabaiana - SE, no Banco Falares Sergipanos e no Banco Fala Natal.

O Banco Falares Sergipanos, criado sem "abrir mão da comparabilidade com os bancos de dados já constituídos" (p. 934) "segue duas linhas de coleta – a de estratificação homogeneizada e a de comunidades de prática" (p. 935). Em Freitag (2013, p. 160-162), estão sintetizadas algumas das diretrizes que concorreram para a sua definição: (a) a pesquisa abrange 6 cidades com 40 entrevistas por cidade, prevendo-se 18 comunidades de práticas (religiosas, recreativas e escolares); (b) a seleção de informantes (2 por célula, a princípio) segue o método "bola de neve", isto é, os indivíduos indicam novos participantes da sua rede de amigos e conhecidos para participarem da pesquisa; (c) a estratificação dos informantes segue a maior/menor distribuição percentual da população do Estado de Sergipe por cinco faixas etárias com base no IBGE (até 14 anos; de 15 a 24; de 25 a 39; de 40 a 64; mais de 65 anos) e, de início, não contempla escolaridade; (d) os informantes passam por uma fase de (i) pré-seleção, checagem para levantamento do seu perfil sociocultural e (ii) entrevista sociolinguística com roteiro prévio, prevendo questões/tópicos que induzam tipos textuais quer narrativos, quer argumentativos/explanativos e, ainda, questões sobre a comunidade e sobre avaliação da fala; (e) a amostra de comunidades de práticas, em cada uma das seis cidades, pretende ser representativa de grupos de indivíduos observados em ação nos espaços escolar, de trabalho, religioso, recreativo, por meio de "gravações de longo termo, em intervalo semanal, por um período de seis meses, a fim

de captar nuanças de estilo e adequação de papéis sociopessoais dos participantes, por meio de coletas longitudinais".

Amostras que servem de base para estudos sobre comunidades de práticas em bancos de dados nos moldes do Falares Sergipanos constitui, sem dúvida, um avanço quanto à observação de fenômenos variáveis, podendo, inclusive, esclarecer aspectos analisados em comunidades de fala, como os relacionados à variável sexo/gênero, por exemplo.

3.1.3 Informatização, disponibilização, centralização e compartilhamento de corpora

Além de aspectos mais propriamente metodológicos, há, ainda, a considerar aspectos práticos, que tornam oneroso, em termos quer financeiros, quer cronológicos, a constituição de corpora, que requer, a depender de sua extensão, uma equipe determinada a realizar entrevistas no mais curto prazo e treinada de forma a evitar vieses que ponham em risco os resultados das análises. Nesse sentido, é de fundamental importância que os integrantes das equipes, embora trabalhando de forma complementar, pela inerente especificidade de cada uma das tarefas próprias da organização e informatização de um corpus, participem ativamente de todas as etapas de trabalho ou, quando se trata da utilização de um corpus já constituído, fiquem inteirados dos parâmetros que o fundamentaram. Assim, parece, a cada dia, mais natural compartilhar corpora por conta, seja do interesse em contrastar diferentes dialetos, seja pelo próprio objeto em análise requerer um campo de observação mais amplo.

A discussão sobre informatização, disponibilização, centralização e compartilhamento de corpora vem de longa data. Em 2007, no II Congresso da Associação Internacional de Linguística Portuguesa (AILP), realizado no Rio de Janeiro, foi organizada a mesa redonda Corpora do Português: formação e políticas de disponibilização, que contou com a participação de Maria Fernanda Bacelar do Nascimento (Universidade de Lisboa), Paulino Vandresen (UFSC), Izete Coelho (UFSC) e Silvia Figueiredo Brandão (UFRJ). Brandão, embora tenha focalizado, em especial, os corpora linguísticos existentes, naquele momento, na Faculdade de Letras da UFRJ, tentou esboçar uma cronologia desses debates, a maior parte deles registrada em Boletins da ABRALIN. O tema, inclusive, já tinha sido debatido, em 2001, em Lisboa, numa outra mesa redonda[8] também no âmbito da AILP, por ocasião do primeiro congresso dessa associação.

[8] Da mesa participaram Fernanda Bacelar do Nascimento (Universidade de Lisboa), Ataliba de Castilho (Universidade de São Paulo) e Perpétua Gonçalves (Universidade Eduardo Mondlane).

Brandão observou que, a partir da década de 1980, algum tempo após a criação e início de desenvolvimento, na década de 1970, de projetos de pesquisa linguística que dependiam da formação de bancos de dados de maior porte, vem, intermitentemente, à discussão a necessidade de se criarem parâmetros e políticas que permitam o compartilhamento de dados e, consequentemente, o incentivo à realização, entre outras, de pesquisas de caráter comparativo.

No histórico apresentado e publicado em 2008 (p. 65-67), ela demonstra que o tema vem sendo discutido há, pelo menos, cerca de 36 anos, sem que se tenha chegado a um consenso sobre alguns tópicos. Os debates se deram pela primeira vez, ao que tudo indica, em 1984, na 36ª Reunião da SBPC, na mesa redonda Problemas de Sociolinguística.[9] Organizada pela ABRALIN, girou em torno de três eixos: (a) reflexões de natureza ética, (b) sugestões no sentido de tornar comparáveis os dados recolhidos por diferentes grupos e delimitar temas de interesse de cada um deles; (c) propostas concretas de levantamento de projetos que contassem com amostras, sendo inclusive apresentado por Sebastião Vôtre um esboço de ficha de dados que permitisse traçar o perfil desses corpora.

Os debates seriam retomados quase 10 anos depois: (a) em 1993, em Campinas, no Seminário sobre a Informatização de Acervos de Língua Portuguesa; (b) em março de 1994, na "Oficina de trabalho sobre Programas de Análise e Tratamento de Textos", em que se fez a demonstração de softwares para tratamento de dados linguísticos[10] (CASTILHO, A.; SILVA, G. M. O.; LUCCHESI, D., 1995, p. 147-148); (c) em julho de 1994, na 46ª Reunião Anual da SBPC, na Universidade Federal do Espírito Santo, no Encontro sobre "Informatização de acervos de língua portuguesa", em que Ataliba T. de Castilho, Giselle Machline de Oliveira e Silva e Dante Lucchesi, após sintetizarem os principais tópicos da referida mesa-redonda e mencionarem algumas iniciativas no sentido de viabilizar o compartilhamento e informatização das amostras, apresentaram um levantamento, ainda que parcial, dos corpora existentes, feito com base em ficha preparada por Rodolfo Ilari e enviada aos grupos de pesquisa pela ABRALIN (CASTILHO, A.; SILVA, G. M. O.; LUCCHESI, D., 1995, p. 148-152).

Voltando à reunião de 1984, no que parece constituir o primeiro debate público sobre corpora linguísticos, o primeiro eixo de discussão dizia respeito a questões de natureza ética.

---

[9] Participaram da mesa Miriam Lemle (UFRJ), Sebastião Vôtre (UFRJ), Claiz Passos (UFBA) e Fernando Tarallo (UNICAMP).

[10] Entre eles, Notebuilder e Wordcruncher; Microconcord e Wordlist; Folio Views; Varbrul, STABLEX. Tact; The Etnograph.

Segundo Pereira e Cardoso (2013, p. 72), no Brasil, remonta a 1988 "a primeira regulamentação referente a pesquisas com seres humanos embasada em documentos e discussões internacionais e focada principalmente na questão biomédica e em pesquisas ligadas a saúde". Segundo as autoras, tal regulamentação teve pouca adesão e a questão só viria a ser retomada em 1995, redundando na Resolução 196/96 do Conselho Nacional de Saúde, que estabelecia diretrizes relativas a pesquisas com seres humanos e criava a Comissão Nacional de Ética em Pesquisa – CONEP. Esse documento "passou por complementações relacionadas a assuntos específicos, ligados a populações indígenas, genética, reprodução humana, dentre outras" (p. 72). Até o final de 2011, os projetos eram enviados a um comitê de ética, vinculado ou não à instituição a que o pesquisador pertencia, e, a partir de janeiro de 2012, encaminhados pela Plataforma Brasil.

Em 1988, ano em que se começa a discutir a questão em nosso país, já se havia iniciado a constituição de bancos de dados de alguns dos grandes projetos que ainda hoje têm continuidade, como o do NURC, em 1973, e a Amostra Censo da Variação Linguística no Rio de Janeiro, de 1976-1977.

Naquela época, sem que houvesse legislação específica sobre procedimentos éticos, as entrevistas eram realizadas com o consentimento oral dos participantes, que tinham seus dados econômico-sociais registrados nas então chamadas fichas dos informantes, cujos dados eram devidamente resguardados, sendo de domínio apenas dos pesquisadores, que atribuíam um código (ou um nome fictício) a cada um deles.

Ao se organizar, por exemplo, na década de 1980, o corpus do Projeto APERJ, continha também elocuções livres que serviram de base a várias análises sociolinguísticas com foco em áreas rurais do Norte e Noroeste fluminenses. O procedimento ético implicava explicar, previamente, que se tinha como objetivo principal recolher a linguagem da pesca e dos pescadores e, ao final da entrevista, perguntar ao informante se queria escutar o seu depoimento, como outra forma de concordância ou não de participação na pesquisa. Era, ainda, praxe, em cada área, visitar as sedes das colônias de pesca e entrar em contato com os líderes comunitários, que acabavam por indicar os pescadores mais representativos, isto é, os indivíduos que se dedicavam precipuamente à pesca.

É fundamental, quando se adentra uma pequena comunidade, ser sensível às normas que regem o convívio entre seus membros e destes com a de forasteiros. Um acontecimento, um detalhe, mesmo involuntário, pode trazer problemas no momento da recolha de dados. Um caso que serve de exemplo foram as entrevistas feitas pela autora deste texto em São Benedito da Lagoa de Cima, no

Município de Campos. Nas duas primeiras vezes que lá foi, tudo correu muito bem, mas, na terceira, nenhum pescador, a princípio, queria dar entrevistas. Só descobriu o motivo, por acaso, quando um dos informantes lhe disse que dois dos entrevistados por ela (inclusive um dos mais jovens e aparentemente muito saudável) tinha morrido de repente, o que a tornou uma pessoa, naquele momento, indesejável. Levou certo tempo, mas readquiriu a confiança do grupo. Dinah Callou, em conversas informais, mencionou que alguns dos potenciais primeiros informantes do NURC, por vezes hesitavam em dar entrevistas, em função do momento político atravessado pelo país nos anos 1970.

Outro dos tópicos de discussão, incluído na rubrica ética no encontro de 1984, dizia respeito à relação entre pesquisadores em diferentes estágios de formação num grupo de pesquisa. É premissa básica que todos os discentes envolvidos na organização de corpora partilhem dos mesmos princípios éticos, sendo fundamental que os coordenadores de projetos promovam atividades com mestrandos, doutorandos ou graduandos de Iniciação Científica de modo a não só lhes transmitir procedimentos metodológicos e éticos, mas também aproveitar suas sugestões. Nos idos da década de 1980, ainda no âmbito do APERJ, os discentes recebiam treinamento para a pesquisa de campo, iam para as áreas de pesquisa, participavam da discussão das normas de transcrição grafemática das entrevistas, normalmente realizada por eles e revista pelos pesquisadores responsáveis pelo projeto.

Hoje, parece já ser consenso não só solicitar ao informante a assinatura de um documento em que explicite sua concordância em participar da pesquisa, mas também, previamente, submeter a uma Comissão de Ética, via Plataforma Brasil, os parâmetros que a nortearão. É essencial, no entanto, que haja em todas as universidades comitês de ética específicos para a área da Linguística.

Certamente, outras questões éticas se imporiam no que concerne à centralização e ao compartilhamento de corpora, como as questões formuladas por Raquel Freitag aos participantes do Simpósio Gestão de dados linguísticos.[11] Foram cinco os questionamentos: (a) como atender aos princípios de ciência aberta quanto ao armazenamento, reuso e autoria do conjunto de dados linguísticos? (b) como lidar com a questão entre transparência na ciência e o sigilo dos participantes? (c) quais as ferramentas mais adequadas para a vitalidade dos conjuntos de dados linguísticos? (d) que ferramentas permitem melhor

[11] Trata-se do simpósio Gestão *de dados linguísticos*, que teve lugar em 21 de julho de 2020 como parte do evento Abralin ao vivo - *Linguists online.* Disponível em: https://aovivo.abralin.org/lives/gestao-de-dados-linguisticos/.

armazenamento e um sistema de interface para consulta e pesquisa? (e) como ficam os grupos minoritários e variedades sub-representadas?

Não há dúvida de que bancos de dados linguísticos, sobretudo os organizados em instituições públicas, devem ser disponibilizados não só para a comunidade científica da área, mas também para a sociedade em geral, pelo seu valor intrínseco e pelo aporte de recursos financeiros que os viabilizam. Além disso, um corpus linguístico pode ser útil a historiadores, antropólogos, sociólogos, historiadores, geógrafos, entre outros.

Em 1993 e em 1994, nos mencionados Seminário e Oficina, já se discutia a necessidade de proceder ao tratamento dos dados por meio dos softwares então existentes. De lá para cá, os recursos de informática avançaram exponencialmente, havendo uma série de opções, algumas, inclusive, aparentemente de fácil manejo, como é o caso dos aplicativos EXMARaLDA e do ELAN, que possibilitam o alinhamento das transcrições com os arquivos de som, permitindo, ainda, que áudio, transcrições e metadados sejam pesquisáveis. Oushiro (2014), inclusive, procura mostrar o caráter amigável do ELAN.

Exemplo bem-sucedido de corpus eletrônico anotado, é o Corpus Histórico do Português Tycho Brahe,[12] idealizado por Charlote Galves e composto de 76 textos em português escritos por autores nascidos entre 1380 e 1881, 44 deles com anotação morfológica e 27, com anotação sintática.

Alguns dos corpora mais longevos já disponibilizam seus dados online, embora ainda de forma convencional, como é o caso do NURC-RJ e do VAL-PB, este com entrevistas realizadas em 1993, em 2015 e 2018. No entanto, como informa Oliveira Jr. (2012), em 2012, por meio de Chamada Universal do CNPq, obtiveram-se os recursos financeiros que permitiram a implantação do NURC-Recife Digital,[13] com corpus anotado e que serviria de modelo para as demais quatro cidades.

Apesar de todos os esforços, a manutenção de sites é um problema recorrente não só pela falta de recursos financeiros regulares, mas também por não se contar com a assessoria permanente de uma equipe (informatas, transcritores de entrevistas, revisores, anotadores de corpora) para as necessárias atualização e manutenção. Grande parte dos bancos de dados está alocada em servidores de instituições públicas e, portanto, na dependência de condições técnicas muito variáveis. No site CORPORAPORT, que visa a divulgar

[12] Cf. http://www.tycho.iel.unicamp.br/corpus/.

[13] Cf. https://fale.ufal.br/projeto/nurcdigital/.

projetos de pesquisa de suas idealizadoras e disponibilizar os corpora que têm servido de base às suas pesquisas e às de seus orientandos, constantemente enfrentam-se dificuldades nesse sentido.

A centralização, em nível nacional, de bancos de dados linguísticos, sem dúvida, seria um grande passo para a preservação e difusão de um acervo linguístico-cultural de valor inestimável, mas constituiria um megaprojeto, que dependeria, entre outros fatores, da constituição de uma equipe multidisciplinar que indicasse se haveria parâmetros de digitalização e anotação comuns aos diferentes bancos de dados ou se eles seriam incorporados no seu formato original. Isso implicaria novas questões de natureza metodológica, a obtenção de recursos financeiros, decisões sobre a representatividade dos corpora, isto é, sobre que bancos de dados seriam reunidos num repositório central e quais os critérios que presidiriam a seu acesso e sua curadoria.

4 BREVES CONSIDERAÇÕES FINAIS

Do que aqui foi exposto, deduz-se que, nos últimos 70 anos, com base na metodologia da Sociolinguística Variacionista e da Dialetologia, muito já se realizou no sentido de conhecer os fenômenos variáveis que atuam no Português do Brasil.

Há, no entanto, muitos desafios para a ampliação desse conhecimento, entre outros: (a) criar novos bancos de dados formatados segundo a metodologia mais adequada para descrever, por exemplo, comunidades que vivem em relativo isolamento, sejam comunidades quilombolas, indígenas, rurais, periféricas aos grandes centros urbanos, para tanto definindo o que se considera isolamento nos dias atuais (geográfico, cultural, econômico, tecnológico); (b) desenvolver metodologia adequada ao estudo de redes sociais e comunidades de prática; (c) ir realimentando os bancos de dados existentes, por exemplo, com o acréscimo de gravações secretas, com entrevistas por meio de aplicativos; (d) buscar métodos que permitam correlacionar resultados obtidos em estudos sobre comunidades de fala e comunidades de prática; (e) encontrar métodos que permitam a comparabilidade quantitativa e qualitativa dos resultados obtidos nas análises, de modo a chegar a generalizações quanto a determinadas variáveis linguísticas, por meio de recursos estatísticos de meta-análise, como vem sendo proposto por Raquel Freitag, tendo em vista que os corpora em que se baseiam análises variacionistas apresentam diferenças quanto ao perfil dos informantes, ao seu número por célula, às variáveis consideradas; (f) proceder à anotação dos

corpora e divulgá-los na web, primeiro passo para a criação de um repositório central de bancos de dados, acessível a todos os cidadãos.

Desafios não faltam, mas também não falta empenho por parte dos pesquisadores no sentido de aprimorar métodos de análise e, por parte da ABRALIN, no sentido de incentivar o compartilhamento de corpora, que é também uma forma de incentivar novas pesquisas.

Johanson (1991, p. 313), em texto intitulado "Times change, and so do corpora", ao especular sobre o futuro dos corpora, faz a seguinte observação, aqui endossada:

> Apesar das grandes mudanças operadas, em menos de três décadas, desde o primeiro corpus concebido para uso em computador, há um aspecto em que o papel do corpus na pesquisa linguística não mudou. O corpus continua sendo uma das ferramentas do linguista, a ser usada junto com a introspecção e técnicas de dedução. Linguistas criteriosos, assim como artífices experientes, afiam suas ferramentas e reconhecem seus usos apropriados.[14] (JOHANSON, 1991, p. 313).

REFERÊNCIAS

ALMEIDA, Fabiana da Silva Campos. Micro Atlas Fonético do Estado do Rio de Janeiro: uma contribuição para o conhecimento dos falares fluminenses. 2 v. Tese (Doutorado em Letras Vernáculas) - Faculdade de Letras, Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2008.

ALUÍSIO, Sandra Maria; ALMEIDA, Gladis Maria de Barcellos. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. Calidoscópio, v. 4, n. 3, p. 156-178, set./dez.2006.

AMARAL, Amadeu. O dialeto caipira. 3. ed. São Paulo: Hucitec; Secretaria de Cultura, Ciência e Tecnologia, 1976 [1920].

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi. Corpus linguistics – Investigating language structure and use. Cambridge: University Press, 1998.

BORTONI-RICARDO, Stela Maris. Educação em língua materna: a sociolinguística na sala de aula. São Paulo: Parábola Editorial, 2004.

[14] "In spite of the great changes in the less than three decades since the first computer corpus, there is one way in which the role of the corpus in linguistic research has not changed. The corpus remains one of the linguist´s tools, to be used together with introspection and elicitation techniques. Wise linguists, like experienced craftsmen, sharpen their tools and recognize their appropriate uses." (Tradução da autora).

BRANDÃO, Silvia Figueiredo. Corpora linguísticos no Rio de Janeiro In: GONÇALVES, Carlos Alexandre Victório; ALMEIDA, Maria Lúcia Leitão de (org.). Língua Portuguesa: identidade, difusão e variabilidade. Rio de Janeiro: AILP/UFRJ, 2008, v. 1, p. 65-73. Disponível em: http://www.ailp-edu.org/download_livro_I.htm. Acesso em: 10 set. 2020.

BRANDÃO, Silvia. Figueiredo. Réalité sociolinguistique brésilienne et géolinguistique pluridimensionnelle. *In:* CARRILHO, Ernestina; MAGRO, Catarina; ÁLVAREZ, Xosé. Current Approaches to Limits and Areas in Dialectology. Cambridge: Cambridge Schollars Publishing, p. 3-26, 2013.

BRANDÃO, Silvia Figueiredo. Pour une approche géo-sociolinguistique de la réalité linguistique brésilienne. Géolinguistique, Grenoble, n. 15, p. 191-214, 2015.

CASTILHO, Ataliba T.; SILVA, Giselle Machline; LUCCHESI, Dante. Informatização de acervos da língua portuguesa. Boletim da ABRALIN, v. 17, p. 143-154, jul. 1995.

DUCROT, Oswald; TODOROV, Tzevetan. Dicionário enciclopédico das ciências da linguagem. 3. ed. São Paulo: Perspectiva, 2001.

ECKERT, Penelope. Three waves of Variation Study: the emergency of meaning in the study of Variation. Annual Review of Antropology, v. 41, p. 87-100, 2012.

FREITAG, Raquel Meister Ko.; MARTINS, Marco Antônio; TAVARES, Maria Alice. Banco de dados sociolinguísticos do português brasileiro e estudos da terceira onda: potencialidades e limitações. Alfa, São Paulo, v. 56, n. 3, p. 917-944, 2012.

FREITAG, Raquel Meister Ko. Banco de dados Falares Sergipanos: falares sergipanos database. Working Papers in Linguística, Florianópolis, v. 13, n. 2, p. 156-164, abr.-jul., 2013.

FREITAG, Raquel Meister Ko. Desafios teórico-metodológicos da sociolinguística variacionista. In: PARREIRA, M. C.; CAVALARI, S. M. S.; ABREU-TARDELLI, L.; NADIN, O. L.; COSTA, D. S. Pesquisas em Linguística no século XXI: perspectivas e desafios teóricos-metodológicos. São Paulo: Cultura Acadêmica, p. 29-43, 2015. Disponível em: https://www.fclar.unesp.br/Home/Instituicao/Administracao/DivisaoTecnicaAcademica/ApoioaoEnsino/LaboratorioEditorial/serie-trilhas-linguisticas-n27.pdf. Acesso em: 18 nov. 2020.

JOHANSSON, Stig. Times change, and so do corpora. *In:* AIJMER, Karin; ALTENBERG, Bengt (org.). English corpus linguistic. Studies in Honour of Jan Svartvik. New York: Longman, 1991. p. 305-314.

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

LABOV, William; ASH, Sharon; BOBERG, Charles. Atlas of North American English: phonetics, phonology and sound change. Berlin: Mouton de Gruyter, 2006.

LEMLE, Miriam. Texto gerador. Boletim da ABRALIN, v. 6, p. 5-11, mai.1984.

LEMLE, Miriam; NARO, Anthony Julius. Competências básicas do português. Relatório final de pesquisa apresentado às instituições patrocinadoras Fundação Movimento Brasileiro de Alfabetização (MOBRAL) e Fundação Ford. Rio de Janeiro, 1977.

LIMA, Luciana Gomes de. 2 v. Atlas Fonético do entorno da Baía de Guanabara-AFeBG. Dissertação (Mestrado em Letras Vernáculas) – Faculdade de Letras, Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2006.

MARGOTTI, Felício. Geolinguística pluridimensional: desafios metodológicos. In: Anais do Encontro do CELSUL, 2008, 9 p. Disponível em: http://www.leffa.pro.br/tela4/Textos/Textos/Anais/CELSUL_VIII/geolinguistica_pluridimensional.pdf. Acesso em: 18 out. 2020.

MENDES, Ronald Belini. A terceira onda da Sociolinguística. *In:* FIORIN, José Luiz. (org.). Novos caminhos da Linguística. São Paulo: Contexto, 2017. p. 103-123.

MOLLICA Maria Cecilia; MELLO, Luciana; LOUREIRO, Fernando; ALÍPIO, Rodrigo. Comunidades rurbanas e conflitos linguísticos. Gragoatá, Niterói, n. 25, p. 63-73, 2. sem. 2008.

NASCENTES, Antenor. O linguajar carioca. 2. ed. Rio de Janeiro: Simões, 1953.

OUSHIRO, Lívia. Transcrição de entrevistas sociolinguísticas com o ELAN. *In:* FREITAG, R. M. K. (org.). Metodologia de coleta em manipulação de dados em sociolinguística. São Paulo: Blucher, 2014. p. 46-50.

PAIVA, Maria da Conceição de; SCHERRE, Maria Marta Pereira. Retrospectiva sociolinguística: contribuições do PEUL. D.E.L.T.A., São Paulo-SP, v. 15, ed. especial, p. 203-230, 1999.

PASSOS, Claiz. Reflexões sobre a profissão de linguista. Boletim da ABRALIN, v. 6, p. 17-26, mai. 1984.

PEREIRA, Lara Rodrigues; CARDOSO, Jaqueline Henrique. Comitês de ética: regulamentando a história oral? Tempos históricos, v. 17, p. 68-82, 2º. sem. 2013.

PEREZ, Aquilino Sanchez. Definicion e historia de los corpus. *In:* PEREZ, A. S.; SARMIENTO, R.; CANTOS, Pascual; SIMÓN, J. (org.). CUMBRE. Corpus lingüístico del español contemporâneo. Fundamentos, metodología y análisis. Madrid: SGEL, 1995.

RADKE, Edgar; THUN, Harald. Radtke Edgar et Thun Harald, Novos caminhos da geolinguística românica. Um balanço. *In:* Radke, E.; THUN, H. (org.). Neue Wege der romanischen Geolinguistik: Akten des Symposiums zur Empirischen Dialektologie (Heilderberg/Mainz, 21-24.10.1991). Kiel: Westensee-Verlag, 1996. p. 25-49.

RODRIGUES, Ângela Cecília de Souza. Concordância verbal: sociolingüística e história do português brasileiro. Fórum Linguístico, Florianópolis, v. 4, n. 1, p. 115-145, jul 2004.

ROSSI, Nelson; ISENSÉE, Dinah Maria; FERREIRA, Carlota. Atlas Prévio dos Falares Baianos. Rio de Janeiro: Instituto Nacional do Livro, 1963.

SARDINHA, Tony Berber. Linguística de corpus, histórico e problemática. D.E.L.T.A, v. 16, n. 2, p. 323-367, São Paulo-SP, 2000.

SCHERRE, Maria Marta Pereira. Breve histórico do Programa de Estudos sobre o Uso da Língua. *In:* SILVA, Giselle Macheline de Oliveira; SCHERRE, Maria Marta Pereira (org.). Padrões sociolinguísticos. Rio de Janeiro: Tempo Brasileiro, 1986. p. 27-50.

SOUZA, Gisela Barcellos. Paisagens rurbanas: a tensão entre práticas rurais e valores urbanos na morfogênese dos espaços públicos de sedes de municípios rurais. Um estudo de caso. Sociedade & Natureza, Uberlândia-MG, v. 21, n. 2, p. 181-192, ago. 2009.

VÔTRE, Sebastião. Para uma política de banco de dados. Boletim da ABRALIN, v. 6, p. 12-16, mai. 1984.