Projeto ALIP (Amostra Linguística do Interior Paulista) e banco de dados *Iboruna*: 10 anos de contribuição com a descrição do português brasileiro

DOI: http://dx.doi.org/10.21165/el.v48i1.2430

Sebastião Carlos Leite Gonçalves[1]

Resumo

Este artigo trata dos aspectos teórico-metodológicos da constituição do Banco de dados *Iboruna*, um banco de dados de médio porte, composto de 152 entrevistas sociolinguísticas (LABOV, 1972) e de 11 interações dialógicas, gravadas secretamente, em contextos de interação social livres (RONCARATI, 1996). Da experiência pioneira no Brasil em disponibilizar integralmente amostras de fala em áudio e as respectivas transcrições, resultaram, no período de uma década, importantes trabalhos de descrição sociofuncionalista de diferentes níveis de análise, do fonético-fonológico ao discursivo-pragmático. A conclusão mais evidente é a de que o avanço da pesquisa linguística baseada no uso da língua depende, cada vez mais, de bancos de dados sistematicamente organizados se tornarem disponíveis, principalmente se tais ferramentas contam com financiamento de recursos públicos.

Palavras-chave: banco de dados; língua falada; português paulista.

1 Universidade Estadual Paulista (UNESP), São José do Rio Preto, São Paulo, Brasil; sebastiao.goncalves@unesp.br; https://orcid.org/0000-0002-1798-729X

Project ALIP (Linguistic Sample from the interior of São Paulo) and *Iboruna* database: 10 years contributing to Brazilian Portuguese description

Abstract

This article deals with the theoretical and methodological aspects of the *Iboruna* database, a medium-sized database composed of 152 sociolinguistic interviews (LABOV, 1972) and 11 dialogic exchanges in free social interaction contexts (RONCARATI, 1996). From the pioneering experience in Brazil to provide audio speech samples and their transcriptions integrally, in the period of a decade, important socio-functionalist descriptions of different levels of analysis, from phonological to discursive-pragmatic, have been produced. The most obvious conclusion is that the advancement of the usage-based linguistic research is increasingly dependent on systematically organized databases becoming available, particularly if such tools are supported by public resources.

Keywords: linguistic database; spoken language; portuguese from São Paulo.

1 Introdução

Completada sua primeira década de existência, o Projeto ALIP (Amostra Linguística do Interior Paulista) e seu banco de dados *Iboruna* (= *rio preto*, em Tupi Guarani) já reúnem contribuições relevantes com a descrição linguística do português brasileiro (PB, daqui em diante). O banco de dados, de médio porte, com pouco mais de 1,5 milhão de palavras, compõe-se de dois tipos de amostras de fala, que registram a variedade do PB falado no interior paulista: (i) uma amostra do censo linguístico de parte da região noroeste do estado de São Paulo, que, nucleada em torno de São José do Rio Preto, se compõe de 152 entrevistas sociolinguísticas; (ii) uma amostra de interação dialógica, constituída de 11 diálogos, envolvendo de dois até cinco informantes, gravados secretamente, em contextos livres de interação social.

O Projeto ALIP conta, hoje, com mais de 300 usuários, do Brasil e do exterior, cadastrados em seu banco de dados, os quais têm acesso livre a toda documentação linguística das duas amostras.

O objetivo deste artigo é mostrar as vantagens de se dispor de bancos de dados sistematicamente organizados para fazer avançar a pesquisa linguística socialmente assentada em contextos reais de uso da língua. Para tanto, inicialmente (seção 2), explicita-se a importância da utilização de banco de dados na pesquisa linguística, apresentando-se dois exemplos de banco de dados de variedades falada e escrita do português paulista (MENDES, 2013; TENANI, 2014), para, na sequência (seção 3), focalizarem-se aspectos teórico-metodológicos da constituição do banco de dados *Iboruna* (GONÇALVES, 2007) e

sua contribuição com trabalhos de descrição sociofuncionalista do português paulista, envolvendo diferentes níveis de análise. As considerações finais reportam-se ao banco de dados constituído e a documentação linguística colocada à disposição de usuários interessados na descrição da língua inserida em seu contexto social.

2 Banco de dados linguísticos e a Linguística de Córpus

Nas últimas décadas do século XX, a chamada *Linguística de Córpus* ganhou impulso e fez avançar estudos de descrição linguística sincrônica e diacrônica assentados em amplas amostras de dados efetivamente atestados, fazendo frente a modelos formais que tomam por base a intuição. De acordo com Biber, Conrad e Reppen (1998), o princípio básico da Linguística de Córpus é o de buscar identificar aquilo que é de uso mais frequente na língua, procurando exceder os limites da simples distinção entre o que faz e o que não faz parte da língua. Para esses autores, mais do que considerar o que é teoricamente possível na língua, pesquisadores que lançam mão de *corpora* em suas investigações estudam a língua real.

Como afirma Sardinha (2002), a Linguística de Córpus, mais do que uma teoria, é uma metodologia baseada no exame minucioso de extensas amostras de dados, ou, nas palavras de Hoey (1997 apud SARDINHA, 2002), é apenas uma rota para a Linguística. Bybee (2016) afirma que, para teorias baseadas no uso, a pesquisa em *corpora* é de extrema relevância para se compreender a amplitude da experiência com o uso da língua, o que significa que modelos linguísticos devem dar conta de detalhe considerável sobre o uso da língua.

Como finalidades precípuas de um banco de dados, podem-se citar: (i) tornar disponível, em meio eletrônico, amostras de fala e/ou de escrita em uma base de dados que agilize pesquisas diversas e torne possível a verificação de hipóteses e postulados teóricos acerca dos efeitos do uso sobre a gramática da língua; (ii) oferecer a pesquisadores interessados bases de dados operacionalizáveis por meio de recursos computacionais; (iii) verificar a produtividade de expressões linguísticas na língua; (iv) possibilitar estudos baseados em extensas amostras de dados efetivamente atestados, de modo a se obter subsídio para a elaboração de gramáticas, dicionários, material para o ensino de língua etc. Por tais finalidades, bancos de dados linguísticos caracterizam-se por constituir um conjunto de amostras de fala e/ou de escrita reunidas em um único lugar de acesso. Coletadas em situações reais de uso da língua, essas amostras podem ser sistematicamente controladas ou não, a depender dos propósitos a que servirá.

Procurando viabilizar ferramentas de busca, banco de dados linguísticos podem comportar anotações específicas e proceder a sua composição de modos diferentes: enquanto bancos com amostras de fala podem conter transcrições das gravações (tipo mais comum), transcrições de gravações e áudios pareados (tipo menos comum),

transcrições, áudio e vídeo pareados (tipo mais raros), além de outras informações sobre as condições de coleta das gravações, bancos com amostras de escritas podem conter textos transcritos de fontes originais ou suas imagens (tipo mais comum), representativos de gêneros discursivos variados de sincronias atuais ou pretéritas. Nesse sentido, ao redor do mundo, são inúmeros os bancos de dados linguísticos de fala e de escrita sistematicamente organizados e disponíveis para pesquisa, dentre os quais podem ser citados, a título de exemplo: o BYU-BNC (*British National Corpus*)[2] e o COCA (*Contemporary Corpus of American English*),[3] para o inglês, o PRESEEA (*Proyecto para el estudio sosiolingüístico del español del España y de América*)[4] e o CORDE (Corpus Diacrónico del Español),[5] para o espanhol, o *Córpus do Português*,[6] o *Córpus Histórico do Português* Thyco Brahe,[7] o CRPC (*Córpus de Referência do Português Contemporâneo*),[8] para o português, dentre tantos outros.

Como exemplos pouco mais detalhados de recentes bancos de dados linguísticos de variedades do português paulista, citem-se o banco de dados do *Projeto SP2010* (MENDES, 2013),[9] o banco de dados *Textus* (TENANI, 2014)[10] e o Banco de dados *Iboruna* (GONÇALVES, 2007), este último a ser tratado na seção seguinte deste artigo, todos eles de acesso livre.

O *Projeto SP2010*, levado a cabo com recursos públicos da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, daqui em diante) (Processo 2011/09278-6), disponibiliza em seu *site* gravações e transcrições de amostras da fala da variedade paulistana do PB, de responsabilidade do GESOL-USP (Grupo de Estudos e Pesquisa em Sociolinguística da USP), com o objetivo de contribuir com a caracterização sociolinguística dessa variedade. Como informam Mendes e Oushiro (2012), de 2008 a 2010, o grupo de pesquisa dedicou-se à organização de seu banco de dados reunindo mais de 100 entrevistas de paulistanos e não paulistanos, de ambos os sexos e orientações sexuais diversas, dos 15 aos 89 anos, com escolaridade variando de Ensino Fundamental incompleto até o superior, de diferentes estratos sociais e moradores de 27 subdistritos e 59 bairros distintos, distribuídos por cinco zonas da capital paulista (Central, Norte, Sul, Leste e Oeste). De caráter exploratório, essas gravações tiveram ainda por objetivo:

2 Disponível em: https://corpus.byu.edu/bnc/

3 Disponível em: https://corpus.byu.edu/coca/

4 Disponível em: http://preseea.linguas.net/Corpus

5 Disponível em: http://corpus.rae.es/cordenet.html

6 Disponível em: https://www.corpusdoportugues.org

7 Disponível em: http://www.tycho.iel.unicamp.br/corpus/

8 Disponível em: http://www.clul.ulisboa.pt/en/10-research/

9 Disponível em: http://projetosp2010.fflch.usp.br

10 Disponível em: http://www.convenios.grupogbd.com/redacoes/Login

> (i) elaborar e aprimorar um roteiro de entrevistas com paulistanos; (ii) elaborar e aprimorar métodos de abordagens a possíveis informantes; (iii) identificar problemas técnicos que podem ocorrer durante a gravação (p.ex., presença de ruídos externos) e metodológicos (p.ex., como conduzir a entrevista de modo que o falante se sinta à vontade), e elaborar soluções para tais problemas; (iv) elaborar critérios para a transcrição das entrevistas; e (v) identificar as variáveis extralinguísticas mais relevantes para a caracterização da fala paulistana. (MENDES; OUSHIRO, 2012, p. 975).

Como se observa no quadro 1, a seguir, a amostra de acesso livre no *site* do *Projeto SP2010* é estratificada de acordo com três variáveis sociais, que constituem parâmetros de busca no banco de dados: *sexo/gênero*, *faixa etária* (de 20 a 34 anos, de 35 a 59 anos, e 60 anos ou mais) e *nível de escolaridade* (até o Ensino Médio, Ensino Superior), com cinco informantes por célula social (p.ex. F2S: sexo feminino, segunda faixa etária, Ensino Superior), distribuídos de modo equânime pelas diferentes regiões e zonas da cidade de São Paulo, não consideradas no tamanho da amostra.

Quadro 1. Grupo de fatores definidores da amostra SP2010 e perfis sociolinguísticos dos informantes

Sexo/gênero (F/M)	**Faixa Etária (1/2/3)**	**Escolaridade (C/S)**	**Perfil Sociolinguístico**
Feminino	20-34 anos	Até Ens. Médio	1. F1C
		Ens. Superior	2. F1S
	35-59 anos	Até Ens. Médio	3. F2C
		Ens. Superior	4. F2S
	60+ anos	Até Ens. Médio	5. F3C
		Ens. Superior	6. F3S
Masculino	20-34 anos	Até Ens. Médio	7. M1C
		Ens. Superior	8. M1S
	35-59 anos	Até Ens. Médio	9. M2C
		Ens. Superior	10. M2S
	60+ anos	Até Ens. Médio	11.M3C
		Ens. Superior	12. M3S
12 perfis x 5 informantes = 60 gravações			

Fonte: Mendes e Oushiro (2012, p. 978)

Como se observa pelos objetivos acima expostos, além das 60 gravações e transcrições,[11] toda documentação linguística disponível na página eletrônica do *Projeto SP2010* (o roteiro de entrevista, os modelos de declaração livre e consentida dos informantes e de fichas do informante e de gravação, o questionário socioeconômico, os mapas de regiões e zonas da cidade de São Paulo, o manual de transcrição e fichas de validação) constitui importante material de referência para a constituição de córpus sociolinguístico. Além disso, o Projeto já conta com vasta produção bibliográfica, incluindo artigos em periódicos, capítulos de livros, teses/dissertações e apresentações em eventos científicos.

O segundo banco de dados, o *Textus*, também constituído sob os auspícios da FAPESP (Processos 2009/14848-6 e 2013/14546-5), comporta 5519 textos manuscritos, produzidos por 662 alunos dos quatros últimos anos do Ensino Fundamental II (6º ao 9º ano) de uma escola pública de periferia da cidade de São José do Rio Preto (SP). O banco organiza-se em torno de uma amostra transversal, com 2759 produções textuais de 470 alunos, coletadas em um mesmo ano letivo, e de uma amostra longitudinal, com 2616 textos, produzidos por 182 alunos ao longo dos quatros anos. Os textos foram coletados durante a realização de oficinas pedagógicas implementadas por meio do projeto de extensão universitária "Desenvolvimento de oficinas pedagógicas de leitura, interpretação e produção textual no ensino fundamental", levado a cabo de 2008 a 2011 (TENANI; LONGHIN-THOMAZI, 2014).

De livre acesso a pesquisadores, o banco de dados *Textus* disponibiliza em sua base os textos manuscritos digitalizados (em imagem JPEG) com as respectivas transcrições ortográficas (em formato DOC), segundo rigorosos critérios de digitação. A base de dados também oferece parâmetros de busca das produções textuais por aluno, sexo/gênero, escola, ano, série, turma, gênero/tipologia textual e tema da proposta da produção escrita (por ex.: Z08_5A_31F_01, que identifica a seguinte produção: escola "Z", Ano 2008, 5ª. série, Turma A, Número do aluno, Sexo Feminino, proposta textual 01). Apenas para exemplificar, o quadro 2 mostra a organização, no banco de dados, dos textos coletados no ano de 2008 de alunos dos 6º e 7º anos.

11 O banco de dados do Projeto *SP2010*, embora disponibilize 60 amostras sociolinguísticas, como mencionado, compõe-se de mais de 100 gravações e encontra-se em contínua expansão.

Quadro 2. Classificação de propostas de produção escrita: tema, gênero e tipologia textual – Quadro sinóptico das propostas (reprodução parcial) [12]

Ano	Proposta	Tema	Gênero	Tipologia Textual
2008 – 6º.	1	Rompimento amoroso	Conto	Narrativa
	2	História pessoal	Cordel	Relato
	3	Poço dos desejos	Relato de experiência	Relato
	4	Internet e uso do MSN	Carta pessoal	Relato
	5	Viagem para outro planeta	Conto	Narrativa
	6	Viagem para Disneylândia	Conto	Narrativa
2008 – 7º.	1	Rompimento amoroso	Carta pessoal	Relato
	2	Experiência perigosa	Relato de experiência	Relato
	3	Jogos Olímpicos de Pequim	Carta pessoal	Relato
	4	Internet e uso do MSN	Carta pessoal	Relato
	5	Viagem para outro planeta	Carta pessoal	Relato
	6	Viagem para Miami	Relato de experiência	Relato

Fonte: Tenani (2014)

Como afirmam Tenani e Longhin-Thomazi (2014), trata-se de um banco de dados inédito que vem subsidiando importantes produções científicas sobre oralidade e letramento, de diferentes níveis, abordando temas como: ortografia convencional, uso de pontuação, aspectos de textualização e articulação de orações, práticas letradas acadêmicas, práticas letradas digitais, dentre outros.

Esses dois exemplos de bancos de dados acima reportados já são suficientes para evidenciar que o acesso a amplos *corpora* permite refinar a apreciação do pesquisador acerca da experiência que o usuário tem com sua língua, permitindo, em última análise, fazer avançar a pesquisa linguística.

12 O quadro 2 se completa com informações de textos produzidos, em 2008, por alunos de 8º e 9º anos, em 2009, por alunos de 7º ano, em 2010, por alunos de 8º ano, e, em 2011, por alunos de 9º ano, a partir de seis ou sete propostas temáticas e de gêneros e tipologia textual diversos.

3 Projeto ALIP e banco de dados *Iboruna*: aspectos teórico-metodológicos

Concebido no interior do Projeto ALIP (Amostra Linguística do Interior Paulista), também subsidiado pela FAPESP (Processo 03/08058-6), o banco de dados *Iboruna* foi constituído, entre os anos de 2003 e 2007, em razão do interesse de um grupo de pesquisadores funcionalistas que, sediado na UNESP de São José do Rio Preto, tem como principal diretriz de suas pesquisas o enfoque da língua usada em seu contexto social. O grupo, filiado ao paradigma funcionalista, identificado especialmente com a Teoria da Gramática Discursivo-funcional (HENGEVELD; MACKENZIE 2008), assume como tarefa descrever a linguagem não como um fim em si mesmo, mas como um requisito pragmático da interação verbal. Sob tal perspectiva, a análise linguística envolve um sistema de regras que governa a constituição das expressões linguísticas (regras semânticas, morfossintáticas e fonológicas), e outro, os padrões de interação verbal em que essas expressões são usadas (regras pragmáticas). O primeiro sistema é visto como instrumental com relação aos objetivos e propósitos do segundo sistema. Desse modo, as expressões linguísticas devem ser descritas e explicadas em termos da organização geral estabelecida pelo sistema pragmático de interação verbal.

Embora haja forte preferência dos membros do grupo por essa vertente funcionalista, outras tendências também são contempladas, como, por exemplo, o funcionalismo da Costa-Oeste americana, hoje denominado *Linguística Cognitiva-funcional*, que tem, nos *Modelos Baseados no Uso* e na *Gramática de Construções*, sua principal referência (BYBEE, 2016; GOLDBERG, 2003), além de tendências sociofuncionalistas, amparadas tanto na Teoria da Variação e Mudança Linguística (WEINREICH; LABOV; HERZOG, 1968) quanto no Paradigma da Gramaticalização (HOPPER; TRAUGOTT, 2003; BYBEE, 2016).

Há dois pontos essenciais que reúnem todas essas tendências funcionalistas: em primeiro lugar, a concepção de linguagem como um "instrumento" de comunicação e de interação social e, em segundo lugar, o estabelecimento de um objeto de estudos baseado no uso real, o que significa não admitir separações entre sistema e uso, dois princípios que fortemente sustentam os trabalhos decorrentes do Projeto ALIP.

Inspirado em grupos de pesquisa brasileiros já bem estruturados,[13] o Projeto ALIP traçou como objetivo primeiro dispor de um banco de dados próprio como recurso fundamental para a consolidação de seu grupo de pesquisadores. Outras motivações fundamentam-se nos seguintes aspectos: (i) a abrangência ainda restrita das descrições dialetais do PB; (ii) a qualidade do material disponível em que tais descrições se embasam; (iii) a validade sincrônica de grande parte das amostras disponíveis, algumas com mais

13 Citem-se, apenas a título de exemplo, o grupo do *PEUL* (Programa de Estudos sobre o Uso da Língua), sediado na UFRJ, e o *Projeto da Gramática do Português culto falado no Brasil*.

de 30 anos, desde sua coleta; (iv) a qualidade acústica das gravações, todas em meios analógicos; (v) o difícil acesso às gravações originais. Além dessas, outra motivação foi a de disponibilizar a pesquisadores um banco de dados anotado[14] com amostras de fala representativa do dialeto do interior paulista, em razão de este ser ainda pouco conhecido em bases científicas, iniciativa que marcou o ineditismo do Projeto, que guardou a preocupação em captar o máximo possível do dinamismo linguístico do PB usado no interior paulista.

3.1 Banco de dados *Iboruna*: 10 anos de contribuição com a descrição do português paulista

O banco de dados *Iboruna*[15] foi idealizado para comportar dois tipos de amostras de fala: uma primeira, tecnicamente denominada *Amostra Censo* ou *Amostra Comunidade* (AC, daqui em diante), segue os preceitos da Sociolinguística, e uma segunda, denominada *Amostra de Interação* (AI, daqui em diante), sem qualquer controle de variáveis sociais, registra interações dialogais gravadas secretamente. Enquanto a primeira é mais propícia a estudos sociolinguísticos, a segunda destina-se a estudos na interface gramática/discurso, uma vez que, sob tal abordagem, concebe-se que a codificação linguística é decisão que decorre de um modelo de interação verbal construído na interlocução. No que se refere aos aspectos éticos da pesquisa, todos os informantes que forneceram dados linguísticos ao Projeto expressaram sua concordância em participar da pesquisa, cujos objetivos lhes foram claramente explicitados.

3.1.1 A Amostra Censo ou Amostra Comunidade

Para a constituição da AC, elegeram-se as variáveis sociais comprovadamente relevantes nos estudos sociolinguísticos (LABOV, 1972). O censo linguístico na região noroeste do estado de São Paulo, nucleada pela cidade de São José do Rio Preto (SJP), englobou seis cidades que lhe fazem fronteira: Bady Bassit (BAD), 12 km ao sul, Cedral (CED), 14 km ao sul, Guapiaçu (GUA), 16 km ao leste, Ipiguá (IPI), 18 km ao norte, Mirassol (MIR), 14 km a oeste, e Onda Verde (OND), 25 km ao norte.

Para conferir representatividade ao censo linguístico, controlaram-se as variáveis elencadas no quadro 3, dado a seguir, que definem a amostra composta de 152 perfis sociais, não incluída em seu dimensionamento a área geográfica do informante, cuja

14 Um banco de dados anotado disponibiliza arquivos sonoros e suas respectivas transcrições ortográficas, com notações pertinentes para o entendimento dos arquivos sonoros.

15 Pretendeu-se atribuir o nome *Iboruna* (= Rio Preto, em tupi-guarani) à cidade de São José do Rio Preto em seu cinquentenário. A intervenção do episcopado local não só impediu a mudança como conquistou definitivamente a denominação primitiva, São José do Rio Preto, reduzida a Rio Preto de 1906 a 1944.

definição segue o método de distribuição aleatória (SILVA, 1996, 2003), uma vez que considerá-la como variável estratificada faria crescer consideravelmente o tamanho da amostra. Assim, distribuíram-se os 152 perfis sociais proporcionalmente à densidade populacional das cidades, segundo o Censo de 2000 (IBGE, 2000), a saber: Bady Bassit: 11.475 habitantes (4 informantes); Cedral: 6.690 habitantes (2 informantes); Guapiaçu: 14.049 habitantes (5 informantes); Ipiguá: 3.461 habitantes (1 informante); Mirassol: 48.233 habitantes (16 informantes); Onda Verde: 5.407 habitantes (2 informantes); São José do Rio Preto: 357.705 habitantes (122 informantes). O método de distribuição aleatória consistiu nos seguintes procedimentos: (i) em uma primeira urna, colocaram-se todos os perfis sociais, numerados de 1 a 152 (v. quadro 3); (ii) em uma segunda urna, colocaram-se os nomes das sete cidades; (iii) por meio de escolha ao acaso, retiravam-se de cada uma das urnas um perfil social e um nome de cidade, definindo-se assim a origem geográfica do perfil sorteado; (iv) em seguida, voltava-se o nome da cidade escolhida para a segunda urna, até que fosse atingido seu número de informantes; (v) nova escolha combinada é feita, até ser definida a origem geográfica de todos os informantes. Esse procedimento de escolha aleatória garante a probabilidade de quaisquer dos perfis sociais pertencerem a uma dada cidade.

A inclusão de mais de um informante por célula social também elevaria consideravelmente o tamanho da amostra, dificultando a exequibilidade do projeto. Como já bem demonstraram outros projetos e o próprio Labov (1972), a variação é bastante padronizada e, mesmo não havendo um imenso número de falantes para sua comprovação, a regularidade linguística emerge, autorizando generalizações acerca da língua usada na comunidade de fala, desde que se atente para duas questões importantes: a necessidade de usar técnicas estatisticamente válidas de amostragem e o conhecimento prévio das dimensões relevantes da estratificação, de forma a poder planejar corretamente a amostragem (PAIVA, 1999).

Quadro 3. Variáveis controladas na constituição da Amostra Censo do Banco de dados *Iboruna*[16]

Renda / gênero / Faixa etária / escolaridade		25+ salários mínimos		11- 24 salários mínimos		6-10 salários mínimos		5- salários mínimos	
		Masc	Fem	Masc	Fem	Masc	Fem	Masc	Fem
7 a 15 anos	1º. Ciclo EF	001	002	003	004	005	006	007	008
	2º. Ciclo EF	009	010	011	012	013	014	015	016
	Ens. Médio	017	018	019	020	021	022	023	024
16-25 anos	1º. Ciclo EF	025	026	027	028	029	030	031	032
	2º. Ciclo EF	033	034	035	036	037	038	039	040
	Ens. médio	041	042	043	044	045	046	047	048
	Superior	049	050	051	052	053	054	055	056
26-35 anos	1º. Ciclo EF	057	058	059	060	061	062	063	064
	2º. Ciclo EF	065	066	067	068	069	070	071	072
	Ens. médio	073	074	075	076	077	078	079	080
	Superior	081	082	083	084	085	086	087	088
36-55 anos	1º. Ciclo EF	089	090	091	092	093	094	095	096
	2º. Ciclo EF	097	098	099	100	101	102	103	104
	Ens. médio	105	106	107	108	109	110	111	112
	Superior	113	114	115	116	117	118	119	120
55+ anos	1º. Ciclo EF	121	122	123	124	125	126	127	128
	2º. Ciclo EF	129	130	131	132	133	134	135	136
	Ens. médio	137	138	139	140	141	142	143	144
	Superior	145	146	147	148	149	150	151	152
Legenda			BAD		OND		IPI		GUA
			CED		MIR		SJP		

Fonte: Gonçalves (2007)

16 O primeiro nível de *faixa etária* representa a fase em que padrões linguísticos estão ainda em fixação e, do segundo em diante, a pressão social sobre a linguagem do indivíduo (SILVA, 1996); a segmentação de *escolaridade*, embora pedagogicamente extinta nos dois primeiros ciclos, preserva diferenças em termos de currículo e metodologia de ensino e representa a divisão escolar em vigor no tempo em que a maioria dos informantes se enquadrava; dada a dificuldade de definição de classe social por indicadores diversos, para a variável *renda familiar*, adotou-se apenas o indicador *salário mínimo*.

As entrevistas da AC foram direcionadas para obtenção de cinco tipos de textos de cada informante, com base em roteiro de entrevista previamente elaborado (VOTRE; OLIVEIRA, 1995): (i) *narrativa de experiência pessoal*, envolvendo relatos de fato pessoal alegre ou triste; (ii) *narrativa recontada*, com reprodução de fato alegre ou triste ocorrido com outrem, sem envolvimento do informante; (iii) *texto descritivo*, baseado em descrição de local; (iv) *relato de procedimentos*, baseado em experiências que exijam procedimentos ordenados; (v) *relato de opinião*, abordando temáticas variadas (escola, família, religião, política etc.).

Resumidamente, a preparação definitiva de AC para o banco de dados compreendeu: (i) preparação dos entrevistadores acerca de questões teórico-metodológicas; (ii) fase piloto de coleta, com avaliação de uma entrevista gravada; (iii) adaptação do roteiro de entrevistas ao perfil sociocultural do informante; (iv) coleta definitiva; (v) validação das gravações; (vi) elaboração de relatório de coleta (diário de campo e ficha social do informante); (vii) transcrição das gravações; (viii) validação das transcrições; (ix) armazenamento eletrônico das gravações e transcrições das amostras no *site* do banco de dados do Projeto.

Somente para efeito de comparação em relação ao tamanho da amostragem de AC, seguem listados, no quadro 4, dado a seguir, alguns bancos de dados de outros projetos conhecidos. Desse quadro, mais do que o tempo de constituição das amostras, interessa destacar, na comparação com o banco de dados do Projeto ALIP, o tamanho e a densidade demográfica das respectivas regiões onde o censo se realizou.

Quadro 4. Quadro comparativo de amostras de fala de diferentes projetos

Projeto	**Abrangência**	**Total de informantes**	**Variáveis controladas**	**Observações**
VALPB (HORA; PEDROSA, 2001)	Estado da Paraíba	60	Sexo, faixa etária e escolaridade	Iniciado em 1993 (não consta a data da conclusão).
Discurso & Gramática (VOTRE; OLIVEIRA, 1995)	Cidade do Rio de Janeiro	96	Sexo, faixa etária e escolaridade	Iniciado em 1991 e disponibilizado em 1995.
PEUL (SCHERRE, 1996)	Cidade do Rio de Janeiro	64 (48 + 16 ampliados)	Sexo, faixa etária e escolaridade	Iniciado em 1980 e concluído em 1983.
VARSUL (VANDRESSEN, 1995)	Região Sul (12 áreas)	288 (24 por área)	Sexo, faixa etária e escolaridade	Iniciado em 1991 (até 1995, não concluído)

Fonte: Elaboração própria.

3.1.2 A Amostra de interação

As gravações de AI, dirigidas para a obtenção de conversações coloquiais e distensas do dia a dia, seguiu metodologia de Roncarati (1996), pioneira em montar um banco de dados interacional no Brasil. Como forma de se garantir a naturalidade do contexto interacional, as gravações foram feitas de modo secreto, dando-se preferência a contextos sociais livres, mas que pudessem manter a qualidade sonora da gravação, requisito que exigiu dos documentadores a posse constante do gravador. Diante dessa estratégia, os perfis sociais para a composição de AI passaram a ser livres, sem o controle de qualquer variável social.[17] Pautadas por essas orientações, foram coletadas 11 amostras de interação, cujas descrições são dadas no quadro 5.

Quadro 5. Informações sobre as interações dialógicas de AI

Identificação da Amostra	**Contexto de interação**
AI-001-CAS	Conversa entre dois homens e três mulheres em ambiente familiar.
AI-002-GIL	Diálogo entre duas amigas vizinhas, no portão da casa de uma delas.
AI-003-ILHA	Diálogo entre tia e sobrinha, em ambiente familiar.
AI-004-OND	Diálogo entre duas irmãs, em ambiente familiar.
AI-005-CAS	Diálogo entre duas estudantes, em ambiente universitário.
AI-006-MAR	Conversa entre quatro mulheres em ambiente familiar.
AI-007-FER	Diálogo entre marido e esposa em ambiente familiar.
AI-008-CAM	Conversa entre três estudantes, masculinos, em ambiente universitário.
AI-009-CAS	Diálogo entre uma cliente e advogado, em escritório de advocacia.
AI-010-CAS	Discussão entre dois advogados sobre peça jurídica.
AI-011-CAS	Diálogo entre um casal de namorados, em ambiente familiar.
11 amostras, com 28 informantes (10 homens e 18 mulheres)	

Fonte: Gonçalves (2007)

17 Mantendo a ética da pesquisa, ao final da gravação, dava-se ciência aos envolvidos da coleta e de seus objetivos. Havendo concordância, preenchia-se a ficha social e colhia-se a declaração de participação livre e consentida; havendo recusa, apagava-se imediatamente a gravação diante dos informantes.

A preparação definitiva de AI para o banco de dados envolveu: (i) validação das gravações; (ii) elaboração do relatório de coleta (diário de campo e ficha social dos informantes); (iii) transcrição das gravações; (iv) validação das transcrições; (v) armazenamento eletrônico das gravações e transcrições das amostras no *site* do banco de dados do Projeto.

3.1.3 Diferenças entre a Amostra Censo e a Amostra de Interação

No quadro abaixo, estão apresentadas as diferenças socioestilísticas entre AC e AI.

Quadro 6. Principais diferenças entre AC e AI

Características	**AC**	**AI**
Tamanho da amostra	152 entrevistas sociolinguísticas	11 interações dialógicas com até 5 informantes, totalizando 28
Perfis sociais dos informantes	Definidos por variáveis sociais	Aleatórios
Participantes da gravação	Documentador e informante (ocasionalmente um terceiro)	Somente informantes (raramente o documentador)
Localização dos informantes	Restrita a perfis sociais da comunidade de fala	Sem restrições
Consentimento dos informantes	Prévio à gravação	Posterior à gravação
Coleta das amostras	Roteirizada previamente e dirigida pelo documentador	Livre, com tópicos definidos pelos participantes da interação
Participação do documentador	Parcialmente ativa; apenas estimulador da entrevista	Passiva
Grau de monitoramento da fala	Maior, minimizando-se o paradoxo do observador (LABOV, 1972)	Menor, tendendo a nulo
Fidelidade ao vernáculo	Menor	Maior

Fonte: Elaboração própria.

3.2 Do sistema de transcrição das amostras de fala

A elaboração de um sistema de transcrição de fala deve nortear-se pelo objetivo básico de procurar transpor o discurso falado, da forma mais fiel possível, para registros gráficos mais permanentes, fidelidade apenas relativa, pois "qualquer notação gráfica do oral é descontínua e dissociativa" (PAIVA, 2003, p. 135). Explicitada a natureza da transcrição, é necessário delimitar e justificar seu grau de detalhamento, cujas convenções acabam por influenciar a percepção dos dados linguísticos (EDWARDS; LAMPERT, 1992), tarefa que requer tomada de decisões teoricamente embasadas e que devem estar claramente explicitadas em manual próprio. Essas foram as diretrizes adotadas na elaboração do manual de transcrição do Projeto ALIP, desenvolvido conjuntamente com os transcritores, que tiveram esclarecidos os princípios organizadores do sistema de transcrição a ser adotado. As convenções adotadas para a transcrição das gravações encontram-se explicitadas no quadro 7, dado a seguir, e são aqui brevemente comentadas.

As convenções sobre a *grafia das palavras* são, relativamente, as mais frequentes nos sistemas de transcrição propostos para o PB, o que, talvez se deva a essas convenções estarem baseadas principalmente nas convenções ortográficas da língua escrita. No entanto, alguns aspectos são anotados de modo diferente do que é prescrito pelas regras gramaticais, tal como o emprego de iniciais maiúsculas restrito a nomes próprios. Ainda sob essa categoria, estão arrolados truncamentos e metalinguagem do informante por contemplarem, em certa medida, aspectos da grafia do que é transcrito da língua oral.[18]

Quadro 7. Convenções gerais do manual de transcrição das amostras de fala do Projeto ALIP

Categorizações	**Convenções**	**Observações / Exemplos**
Grafia das palavras		
Nomes próprios[19]	Com iniciais maiúsculas	a festa foi na casa do ***C****arlos*?
Nomes de obras ou palavras estrangeiras	Grafados na língua original e em itálico	ele adorava ouvir ***Purple Rain***
Marcadores discursivos interrogativos	Seguidos de ponto de interrogação	é pra deixar aqui **né?**
Fáticos/interjeições	Seguidos de ponto de exclamação	**ah!** ... que alívio

(continua)

18 *Truncamento* indica a ocorrência de palavras incompletas ou interrupção brusca.

19 Nome próprio que identifica o informante ou pessoa próxima dele foi indicado apenas pela inicial em maiúscula (p.ex.: C. em lugar de Carlos).

Numerais	Grafados por extenso	passaram **trinta e três** alunos
Truncamentos	Marcados por barra	... ele **ca/** casou semana passada
Metalinguagem do informante	Marcada por 'aspas simples'	o **'s'** do carioca é chiado
Aspectos prosódicos		
Silabação	Uso de hífen entre as sílabas	ele disse **–** "**fi-que-a-qui**"
Pausa (longa ou breve)	Marcada por reticências	ele **...** voltou feliz
Ênfase	Indicada por caixa alta	ele almoçou com **ELA**...
Alongamento de sons	Marcado por dois pontos seguidos	cê **a::**cha?
Pergunta	Uso de ponto de interrogação	B., cê pode me contar sua viagem**?**
Aspectos da interação		
Identificação dos participantes	Documentador (Doc.), Informante (Inf) e Interveniente (Int)	
Início de turno	Em letra minúscula	Doc.: **q**uando foi Inf.: **f**az uns dois meses....
Discurso direto	Indicado por aspas duplas e duplo travessão	ela disse **– – "Vamos à festa?" – –** eu respondi **– – "talvez" – –**
Mudança do fluxo discursivo	Indicada por duplo travessão	eu não tinha **– – fique quieto!** ((falando com o cachorro)) **– –** tempo de estudar
Superposição de vozes	Texto entre colchetes, com índice sobrescrito à esquerda do colchete inicial das falas sobrepostas	Inf.1: eu não tinha saído de lá... **[1][e foi então...]** Doc.: **[1][cê tava]** em casa ainda?
Comentários do transcritor		
Hipótese do que se ouviu	Entre parênteses	aí ele **(virô)** e disse
Comentário descritivo	Entre parênteses duplos	((risos)) ((tossiu)) ((ruído))

Fonte: Gonçalves e Tenani (2008, com adaptações)

O segundo conjunto de convenções trata de *alguns aspectos prosódicos*, como pausa, duração (alongamento de vogais e/ou consoantes), ritmo e velocidade de fala (silabação),[20] entoação (somente o padrão da interrogativa direta) e variação de altura e intensidade percebida como 'ênfase'.[21] Sobre cada um desses elementos prosódicos, várias observações poderiam ser feitas, mas comentam-se aqui apenas duas delas. A primeira trata da escolha desses e não de outros elementos prosódicos: a recorrência com que, por exemplo, a pausa é transcrita parece indiciar certa facilidade de percepção (auditiva, em geral), sem entrar em questão sua duração; some-se a isso a função delimitadora de fronteira prosódica (coincidentes ou não com fronteiras sintáticas, por exemplo), outra razão (não explicitada geralmente) que torna quase obrigatória a identificação da pausa. A segunda observação recai sobre o grau de detalhamento dos elementos prosódicos transcritos, que pode ser formulado nos seguintes termos: é suficientemente adequado propor que a seleção dos elementos prosódicos e que seu grau de detalhamento seja baseado na (relativa facilidade de) percepção do transcritor (falante nativo da variedade do português analisado)? A solução adotada inicialmente foi transcrever os elementos prosódicos mais frequentemente anotados nos sistemas de transcrição e com o mesmo grau de refinamento. Um cuidado extra consistiu na realização de treinamento dos transcritores a fim de explicitar técnicas de transcrição de base auditiva, somadas a recursos de *softwares* de análise de fala, como o PRAAT, de maneira que houvesse homogeneização do material transcrito.

A categoria de *aspectos da interação* visa assegurar a identificação dos participantes da interação, além de indicar também a *mudança do fluxo discursivo*, entendida como o momento em que o informante se dirige a um outro interlocutor diferente do documentador, categoria que não deve ser confundida com *mudança ou desvio de sequência temática*, que requereria uma análise prévia do texto oral.[22]

A última categoria, *comentários do transcritor*, é recurso que dá visibilidade à presença do transcritor, embora a transcrição seja permeada por suas escolhas (preferencialmente, orientadas por um sistema que pretende a homogeneização – quer para as hipóteses, quer para os comentários a serem expressos).

---

20 Ritmo e velocidade de fala são elementos prosódicos distintos, embora muito frequentemente confundidos. A rigor, a velocidade varia de modo independente do padrão rítmico da língua. Um enunciado de uma língua de ritmo acentual, como o Português Brasileiro, pode ser realizado em velocidade 'neutra' (em *andante* ou ainda em *alegro*) (MORAES; LEITE, 1993).

21 Nos sistemas de transcrição consultados, a ênfase sempre é indicada, mas nunca explicitamente.

22 Para a categoria *desvio de sequência temática*, duas questões se colocam: (i) o pressuposto teórico adotado e (ii) os critérios para a sua identificação. Atendido (i), resta o desafio de considerar essa categoria como parte do sistema de transcrição, pois suas marcas linguísticas são de natureza diversa (prosódica, morfossintática, léxico-semântica etc.) e não são facilmente apreensíveis. Por essa razão, essa categoria não foi contemplada no sistema de transcrição proposto.

3.3 Das dificuldades na execução de composição do banco de dados

Não se devem negligenciar as dificuldades encontradas na execução do Projeto ALIP, as quais, aqui brevemente sumarizadas, dizem respeito: (i) ao uso de gravadores digitais disponíveis à época no mercado,[23] os quais, de alto custo, necessitavam de *software* específico para a extração e conversão de arquivos em formato *WAVE* ou *MP3*, levando à perda de qualidade da gravação ou mesmo da gravação completa; (ii) ao trabalho de transcrição das gravações: apesar do intenso treinamento da equipe técnica do Projeto para essa tarefa, a homogeneidade na aplicação das convenções estabelecidas em manual esteve longe de ser alcançada; (iii) a dificuldade de localização na comunidade de certos perfis sociais de AC, dada a restrição de o informante ter de ser nascido na região ou ter se tornado residente dela antes dos cinco anos de idade, sem dela poder ter se ausentado por mais de dois anos.

As soluções para esses problemas exigiram medidas que atrasaram em certa medida o cronograma do projeto, tais como: (i) aquisição de gravadores digitais modernos e mais caros; (ii) regravação de entrevistas perdidas por problemas de *software*; (iii) árduo trabalho de revisão das transcrições por uma equipe menor e, por último, por um único revisor; (iv) alargamento, quando extremamente necessário, das fronteiras do critério social que define perfil do informante para facilitar sua localização na comunidade.

3.4 Principais contribuições do Projeto ALIP

Fechando esta seção, registram-se, por último, as contribuições pioneiras do banco dados *Iboruna* com o trabalho de descrição do PB em sua variedade paulista, contribuições cujos autores encontram-se devidamente referenciados em Gonçalves e Rubio (2012).

Da perspectiva variacionista, os fenômenos já investigados contemplam os níveis fonético-fonológico e morfossintático. Mais especificamente, as contribuições incluem resultados para os seguintes fenômenos variáveis: (i) alçamento vocálico em contextos de postônica medial de nomes, como em *c**[o]**zinha* ~ *c**[u]**zinha* e *t**[e]**soura* ~ *t**[i]**soura*, e de verbos, como em *d**[e]**via* ~ *d**[i]**via* e *p**[o]**dia* ~ *p**[u]**dia*; (ii) alçamento e síncope de postônicas mediais, como em *pês.s**[e]**.go* ~ *pês.s**[i]**.go* ~ *pês.go* e *a.bó.**[bo]**.ra* ~ *a.bó.**[bu]**.ra* ~ *a.bó.* ***[bra]***; (iii) redução de gerúndio, como em *canta**[ndo]*** ~ *canta**[no]***, traço marcante da fala paulista interiorana, que alcança percentuais elevadíssimos de aplicação da alternante reduzida. Resultados para fenômenos variáveis de ordem morfossintática incluem: (i) alternância entre futuro sintético e analítico, como em ***vou*** *cantar* ~ *canta**rei***; (ii) expressão de aspecto cursivo por meio de perífrases, como em *andar* ~ *continuar* ~ *ficar* ~ *viver* + gerúndio; (iii) realização de preposições com e sem contração, como em *com a* ~ *cu'a* ~ *c'a*, *para* ~ *pra* ~ *pa* etc.; (iv) expressão de cópula e complementizador em orações matrizes predicativas,

23 Os gravadores utilizados eram das marcas *gama-power* e *power-pack*.

como em *é claro* ***que*** ~ *claro* ***que*** ~ *é claro*; (v) alternância indicativo/subjuntivo em orações complexas, como em *quero que você* ***vai*** ~ *quero que você* ***vá***; (vi) marcação de plural em sintagmas nominais e em contextos de predicativo; concordância verbal de primeira e terceira pessoas do plural; padrões de concordância verbal e de alternância pronominal entre *nós* e *a gente*.

Rotulado, de modo mais amplo, de *sociofuncionalista*, outro conjunto de trabalhos combina as perspectivas da variação e da gramaticalização. Incluem-se nessa vertente: (i) investigação de preposições com verbos de movimento; (ii) a alternância *nós x a gente*; (iii) alçamento de constituintes em construções complexas; (iv) uso de predicações reduzidas encaixadas em predicados avaliativos.

O banco de dados *Iboruna* tem servido ainda ao desenvolvimento de inúmeras pesquisas de descrição do português falado sob a perspectiva funcionalista, não variacionista. Dentre as várias pesquisas já concluídas, citem-se: (i) uso de marcadores discursivos; (ii) gramaticalização de juntivos; (iii) processos de combinação de orações; (iv) expressão de modalidade e de evidencialidade; (v) uso de predicações não-verbais.

Considerações finais

Sediado na UNESP de São José do Rio Preto e financiado integralmente com recursos públicos, o Projeto ALIP (Amostra Linguística do Interior Paulista) permite acesso livre a seu banco de dados, de onde podem ser obtidos, além das transcrições ortográficas das gravações, os respectivos arquivos de som, as fichas sociais dos informantes, as fichas de validação do material sonoro, os diários de campo, o roteiro de entrevistas sociolinguísticas e o manual de transcrição das amostras de fala. Atualmente, são mais de 300 pesquisadores cadastrados como usuários do banco de dados, incluindo pesquisadores do Brasil e do exterior, o que atende satisfatoriamente à expectativa do Projeto de dar acesso livre a interessados em pesquisas baseadas no uso da língua, principalmente quando há investimento de recursos públicos.

Os resultados aqui apresentados, ao mesmo tempo em que servem de instrumento de divulgação do banco de dados *Iboruna*, por ocasião de seus 10 anos de existência, e de debate da metodologia empregada, servem também de guia prático para a execução de projetos semelhantes, dada a dimensão da amostra constituída. De toda essa exposição, a conclusão mais evidente que deve ficar é a de que o avanço da pesquisa linguística baseada no uso da língua depende, cada vez mais, de tornar disponíveis bancos de dados sistematicamente organizados, principalmente se tais ferramentas contam com financiamento de recursos públicos, prática ainda muito tímida no Brasil.

REFERÊNCIAS

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus linguistics*: investigating language structure and use. Cambridge: Cambridge University Press, 1998.

BYBEE, J. *Língua*, *uso e cognição*. Tradução Maria Angélica Furtado da Cunha e Sebastião Carlos Leite Gonçalves. São Paulo: Cortez, 2016.

EDWARDS, J. A.; LAMPERT, M. D. (ed.). *Talking data:* transcription and coding in discourse research. New Jersey: Lawrence Erlbaum Associates, 1992.

GOLDBERG, A. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, v. 7, n. 5, p. 219-224, 2003.

GONÇALVES, S. C. L. *Banco de dados Iboruna*: amostras eletrônicas do português falado no interior paulista. 2007. Disponível em: http://www.iboruna.ibilce.unesp.br. Acesso em: 15 jun. 2018.

GONÇALVES, S. C. L.; RUBIO, C. F. A fala do interior paulista no cenário da sociolinguística brasileira: panorama da concordância verbal e da alternância pronominal. *Alfa*, São Paulo, n. 56, v. 3, p. 1003-1034, 2012.

GONÇALVES, S. C. L.; TENANI, L. E. Problemas teórico-metodológicos na elaboração de um sistema de transcrição de dados interacionais: o caso do Projeto ALIP (Amostra Linguística do Interior Paulista. *Gragoatá*, Niterói, n. 25, p. 165-183, 2008.

HENGEVELD, K.; MACKENZIE, J. L. *Functional discourse grammar:* a typologically-based theory of language structure. Oxford: University Press, 2008.

HOPPER, J.; TRAUGOTT, E. *Grammaticalization*. 2. ed. Cambridge: Cambridge University Press, 2003.

HORA, D.; PEDROSA, J. L. R. (org.). *Projeto variação lingüística no Estado da Paraíba (VALPB)*. João Pessoa: Idéia, 2001.

IBGE. Instituto Brasileiro de Geografia e Estatística. *Censo demográfico 2000*. Brasília, 2000. Disponível em: http://www.ibge.gov.br/home/estatistica/populacao/default_censo_2000.shtm. Acesso em: 20 out. 2002.

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

MENDES, R. B. *Projeto SP2010*: Amostra da fala paulistana. 2013. Disponível em: http://projetosp2010.fflch.usp.br. Acesso em: 15 jun. 2018.

MENDES, R. B.; OUSHIRO, L. O paulistano no mapa sociolinguístico brasileiro. *Alfa*, São Paulo, n. 56, v. 3, p. 973-1001, 2012.

MORAES, J. A.; LEITE, Y. F. Ritmo e velocidade de fala na estratégia do discurso: uma proposta de trabalho. *In*: ILARI, R. (org.). *Gramática do português falado*. Campinas: Editora da UNICAMP, 1993. p. 67-77.

PAIVA, M. C. Transcrição de dados lingüísticos. *In*: MOLLICA, M. C.; BRAGA, M. L. (org.). *Introdução à Sociolingüística*. São Paulo: Contexto, 2003. p. 135-143.

PAIVA, M. C. (org.). *Amostras de fala do português falado no Rio de Janeiro*. Rio de Janeiro: UFRJ, 1999.

RONCARATI, C. N. (org.). *Bancos de dados interacionais do Programa de Estudos Sobre o Uso da Língua.* Rio de Janeiro: Divisão Gráfica/UFRJ, 1996.

SARDINHA, T. B. Análise multidimensional. *D.E.L.T.A*, São Paulo, v. 16, n. 1, p. 99-127, 2002.

SCHERRE, M. M. P. Breve histórico do Programa de Estudos Sobre o Uso da Língua. *In*: SILVA, G. M. O.; SCHERRE, M. M. P. (org.). *Padrões sociolingüísticos*: análise de fenômenos variáveis do português falado na cidade do Rio de Janeiro. Rio de Janeiro: Tempo Brasileiro, 1996. p. 27-36.

SILVA, G. M. O. Variáveis sociais e perfil do *corpus CENSO. In*: SILVA, G. M. O.; SCHERRE, M. M. P. (org.). *Padrões sociolinguísticos*. Rio de Janeiro: Tempo Brasileiro, 1996. p. 51-81.

SILVA, G. M. O. Coleta de dados. *In*: MOLLICA, M. C.; BRAGA, M. L. (org.). *Introdução à Sociolingüística*: o tratamento da variação. São Paulo: Contexto, 2003. p. 117-134.

TENANI, L. E. *Banco de dados de escrita do ensino Fundamental II*. 2014. Disponível em: http://www.convenios.grupogbd.com/redacoes/Login. Acesso em: 15 jun. 2018.

TENANI, L. E.; LONGHIN-THOMAZI. Oficinas de leitura, interpretação e produção textual no ensino fundamental. *Em Extensão*, Uberlândia, v. 13, n. 1, p. 20-34, jan./jun. 2014.

VANDRESSEN, P. O Projeto Varsul – Avaliação e perspectivas sobre pesquisas do português falado na Região Sul. *In*: ENCONTRO NACIONAL DE LÍNGUA FALADA E ENSINO, 1, 1995, Maceió. *Anais...* Maceió: EdUFAL, 1995. p. 196-221.

VOTRE, S.; OLIVEIRA, M. R. *A Língua falada e escrita na cidade do Rio de Janeiro*: materiais para seu estudo. Rio de Janeiro: UFRJ, 1995.

WEINREICH, U.; LABOV, W.; HERZOG, M. Empirical foundations for a theory of language change. *In*: LEHMAN, W.; MALKIEL, Y. (ed.). *Directions for historical linguistics.* Austin: University of Texas Press, 1968. p. 97-195.