PROJETO ALIP (AMOSTRA LINGÜÍSTICA DO INTERIOR PAULISTA) [*]

Sebastião Carlos Leite GONÇALVES [**] (UNESP)

RESUMO: Este artigo trata da divulgação do projeto ALIP (Amostra Lingüística do Interior Paulista), que, hoje, comporta um banco de dados do português falado na região de São José do Rio Preto, mais precisamente em sete cidades vizinhas situadas na região noroeste do Estado de São Paulo. O banco de dados, denominado *Iboruna* (= *rio preto*, em Tupi), é composto por dois diferentes tipos de amostra: uma amostra do censo lingüístico da região, com o controle rigoroso de variáveis sociais (*sexo*/*gênero*, *idade* e *escolaridade*), e uma amostra de interação dialógica, composta de diálogos gravados secretamente em contextos de interação social livres.

ABSTRACT: This paper leads with divulging the ALIP (Amostra Lingüística do Interior Paulista) Project, which has a data bank with samples from Portuguese spoken in the region of Sao Jose do Rio Preto, more specifically in seven neighbouring cities located in the northwest of Sao Paulo State, Brazil. The data bank, named *Iboruna* (= *black river*, in Tupi), is compounded from two types of sample: the first one comes from the linguistic *census* of the region and controls the social variables *sex/gender*, *age*, *level education*; the second one comes from dialogic interactions, taped secretly from free social interaction contexts.

1. Introdução

O presente artigo tem por objetivo divulgar o projeto ALIP (Amostra Lingüística do Interior Paulista) instituído a partir da proposta de constituição de um banco de dados com amostras de fala, que, hoje, vem servindo a trabalhos de descrição do português brasileiro, na sua variedade falada no interior do Estado de São Paulo. Sediado na UNESP de São José do Rio Preto, o Projeto ALIP coloca o seu banco de dados à disposição dos interessados na descrição lingüística feita a partir de manifestações reais da linguagem inserida no contexto social.

Motivados por essa disposição, apresentamos, neste artigo, um breve histórico da instituição do Projeto ALIP (seção 2), as orientações teóricas e metodológicas que nortearam a constituição do banco de dados (seção 3) e as dificuldades encontradas na sua implantação (seção 3). Esperamos, assim, que, ao mesmo que sirva de instrumento de divulgação, este artigo sirva também de guia prático na execução de projetos semelhantes.

2. O projeto ALIP

2.1. Breve histórico

O projeto ALIP foi uma iniciativa concebida no interior do Grupo de Pesquisa em Gramática Funcional, (GPGF) do IBILCE/UNESP, entre os anos de 2002 e 2003, em razão do interesse dos membros do grupo na "Descrição Funcional do Português Oral e Escrito", linha de pesquisa que tem como principal diretriz o enfoque da língua usada no seu contexto social. O projeto se constituiu sob dos auspícios da Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP (Proc. 03/08058-6).

O GPGF vem se despontando como núcleo de pesquisa emergente, preocupado que está com a conjugação da investigação lingüística de orientação funcionalista, identificada especialmente com a Teoria da Gramática Funcional, escola holandesa, que tem em S. Dik (já falecido), K. Hengeveld e J. L. Mackenzie seus principais mentores. O compromisso principal desse enfoque funcionalista é descrever a linguagem não como um fim em si mesmo, mas como um requisito pragmático da interação verbal (DIK, 1989, p. 2). Sob tal perspectiva, a análise lingüística envolve dois tipos de sistemas de regras: (i) um que governa a constituição das expressões lingüísticas (regras semânticas, sintáticas, morfológicas e fonológicas); (ii) e outro que governa os padrões de interação verbal em que essas expressões lingüísticas são usadas (regras pragmáticas). O sistema em (i) é visto como instrumental com relação aos objetivos e propósitos do sistema em (ii): o requisito básico do paradigma funcional é que as expressões lingüísticas devem ser descritas e explicadas em termos da organização geral estabelecida pelo sistema pragmático de interação verbal.

[*] Agradeço à FAPESP (Proc. 03/08058-6) o financiamento dessa pesquisa.

[**] Universidade Estadual Paulista – UNESP. Instituto de Biociências, Letras e Ciências Exatas – IBILCE. Departamento de Estudos Lingüísticos e Literários. Grupo de Pesquisa em Gramática Funcional. scarlos@ibilce.unesp.br. São José do Rio Preto (SP).

Embora haja uma forte preferência por essa corrente holandesa, a produção científica do GPGF baseia-se em outras tendências funcionalistas, como as desenvolvidas pelos seguintes grupos: (i) os pesquisadores da Califórnia (EUA) de diversas tendências, entre os quais se incluem T. Givón, C. Li, S. Thompson, W. Chafe, P. Hopper, S. DeLancey, J. DuBois entre outros; (ii) os pesquisadores da 'Gramática de Papel e de Referência' (*Role and Reference Grammar*), centrados em Buffalo, New York, em torno de R. Van Valin Jr; (iii) os pesquisadores da tendência funcional-cognitiva de que se destacam as figuras de G. Lakoff, M. Johnson, R. Langacker e E. Sweeteser; os pesquisadores que desenvolvem trabalhos no âmbito do 'Paradigma da Gramaticalização', tais como P. Hopper, B. Heine, C. Lehmann, E. Traugott. Ainda uma tendência sociofuncionalista, amparada na 'Teoria da Variação e Mudança Lingüística', norteia a produção científica dos membros do GPGF.

Há dois pontos essenciais que reúnem todas essas tendências, com as quais o GPGF se mostra particularmente identificado: 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, tal como preconizam em geral as correntes formalistas. São esses dois princípios que dão suporte teórico-metodológico aos trabalhos que hoje são decorrentes do projeto ALIP.

2.2. Motivações

Inspirado na origem e no reconhecimento de importantes grupos de pesquisa brasileiros, hoje já bem estruturados,[1] o objetivo premente do GPGF, ao propor o Projeto ALIP, era o de constituir um banco de dados com amostras de fala, recurso de fundamental importância para a caracterização e a **consolidação de um grupo de pesquisa**. O compromisso teórico com as pesquisas que vínhamos desenvolvendo, até então, nos fazia lançar mão de amostras de falas já sistematizadas, definidas a partir de objetivos pré-estabelecidos, dados os interesses dos grupos que as detinham. Assim é que a diversidade de *corpora* utilizada pelos membros do GPGF em suas pesquisas[2] representava um obstáculo a ser superado, desejo que, muitas vezes, levava seus integrantes a despenderem tempo, energia e recursos na montagem de amostras que vinham a atender interesses de pesquisa mais específicos.[3]

Outras motivações que levaram à proposição do Projeto ALIP encontram sua fundamentação na pouca abrangência das descrições feitas para o português atualmente usado no Brasil e na qualidade do material disponível em que tais descrições se embasam.

Com raríssimas exceções, os trabalhos de descrição do português, que assumiram para si a tarefa de descrição da língua inserida no seu contexto social, ainda não foram capazes de oferecer uma **visão ampla e mais realística do PB**. A preocupação com variáveis específicas recorta a dimensão mais ampla da amostra usada como referência para o trabalho descritivo.[4] Descarta-se, assim, uma gama de dialetos/socioletos, à medida que se encontram representados nessas amostras, por exemplo, apenas dialetos/socioletos provindos das grandes metrópoles do país (v. SILVA, 1996; CASTILHO, 1990; PRETI & URBANO, 1986; CASTILHO & PRETI, 1986, 1987; VOTRE e OLIVEIRA, 1995), de um estrato com elevado nível de escolaridade, pouco representativo da população brasileira, e de faixas etárias acima de 25 anos (v. CASTILHO, 1990; PRETI & URBANO, 1986; CASTILHO & PRETI, 1986, 1987). Cabe, nesse ponto, lembrar que essas restrições, ainda que pautadas por **objetivos cientificamente legítimos**, contribuem também para aumentar o já acentuado estigma em torno do português não-culto, se podemos assim nos referir, e "interiorano", por vezes lembrado somente em pesquisas esparsas e de pouca representatividade, casos em que não se devem incluir os estudos dialetológicos, cujas preocupações mais recentes para a elaboração de Atlas Lingüísticos voltam-se para um espectro mais amplo de fatos lingüísticos (v. OLIVEIRA & ISQUERDO, 2003) .

Outra motivação refere-se à **validade sincrônica** de grande parte das amostras disponíveis, as quais já vão para mais de 30 anos. Esse problema tem sido resolvido pela atualização das amostras de fala com as quais alguns grupos operam, preocupação bem retratada pelos estudos pioneiros do PEUL, que enfocam a mudança lingüística em tempo real de curta duração, caminho possível somente a partir de nova coleta, equivalente àquela levantada no início dos anos de 1980 (v. PAIVA & DUARTE, 2003).

---

[1] Citemos, a título de exemplificação, o grupo de pesquisadores do PEUL (Programa de Estudos sobre o Uso da Língua), sediado na UFRJ (cf. SCHERRE, 1996; SILVA, 1996; PAIVA, 1999), do VARSUL (Variação Lingüística Urbana na Região Sul do País), sediado na UFSC (cf. Lira, 1987), do "Gramática & Discurso", também sediado na UFRJ (cf. VOTRE & OLIVEIRA, 1995) e do PGPF (Projeto da Gramática do Português falado), sediado na UNICAMP (cf. CASTILHO, 1990; CASTILHO & PRETI, 1986).

[2] Nas pesquisas de docentes e de alunos do grupo, os seguintes *corpora* de língua falada já foram usados: *Corpus* mínimo do Projeto da Gramática do Português Falado, *Corpus* do Grupo "Gramática & Discurso", Amostras *Censo* e *Tendência* do PEUL e *Corpus* do NURC/RJ, além de *corpora* específicos de língua escrita.

[3] Um exemplo desse esforço é o trabalho de Castro (2002), que investigou a realização do /r/ retroflexo no dialeto riopretense.

[4] Fazemos aqui também referência aos trabalhos reunidos nos oito volumes da *Gramática do Português Falado*, cuja preocupação é constituir-se gramática de referência, que toma por base o padrão culto da língua.

Além desses "problemas" de ordem metodológica que restringem o registro de uma dimensão mais ampla do português, um outro fator limitante diz respeito à **qualidade acústica das gravações**, todas elas feitas por meio de recursos analógicos, o que, de certa forma, dificulta o estudo de variáveis fonético-fonológicas e prosódicas. Além disso, o acesso às gravações originais nem sempre é possível e as regravações geralmente implicam perda de qualidade sonora.

Essas reflexões apontam, então, para a necessidade de estudos analíticos mais sistemáticos e abrangentes sobre características de variedades do português que privilegiem variáveis relegadas nos estudos de maiores dimensões, e que, certamente, podem vir a complementar estudos descritivos já realizados sobre o PB.

2.3. Objetivos

Com a execução do Projeto ALIP, nosso objetivo foi o de, inicialmente, constituir um banco de dados anotado[5] com amostras de fala sistematicamente controladas por variáveis sociais e representativo, portanto, do dialeto falado no interior paulista, em razão de este ser ainda pouco conhecido, em bases científicas, por seus usuários, e pelos próprios lingüistas. Essa iniciativa marca o ineditismo do projeto, uma vez que não há, até onde sabemos, banco de dados organizado com amostras do português falado no interior do Estado de São Paulo.[6]

Embora inspirados em outros projetos já consolidados (NURC, VARSUL, PEUL, VALPB, LUAL, *Gramática & Discurso* etc), devemos deixar claro que não constituiu prerrogativa do GPGF o interesse por algum fenômeno lingüístico em particular, marca característica das amostras desses outros projetos, a maioria deles constituída com a finalidade precípua de estudos de cunho variacionista (v. LIRA, 1987; SCHERRE, 1996, HORA & PEDROSA, 2001, MOURA, 1995) ou de estudos da iconicidade diagramática (cf. VOTRE & RIOS, 1995, CUNHA, 1998). Por essa razão, servimo-nos apenas dos aspectos teórico-metodológicos desses projetos, conforme será explicitado mais adiante. A concepção do Projeto ALIP, portanto, embasa-se em uma proposta mais ampla e mais aberta, guardando a preocupação de captar o máximo possível do dinamismo lingüístico do PB usado na região de São José do Rio Preto.

Desse modo, a expectativa é a de que esse banco de dados propicie material lingüístico para a descrição do PB e forneça condições para a validação e o desenvolvimento de teorias lingüísticas, representando, assim, importante ferramenta de pesquisa. Como desdobramento desses objetivos mais gerais, as amostras de fala poderão também constituir material subsidiário para atividades de ensino e de pesquisa não necessariamente vinculadas aos interesses sociofuncionalistas do GPGF.

3. O Banco de dados *Iboruna* [7]

O banco de dados *Iboruna* foi idealizado para comportar dois tipos de amostras de fala: uma primeira amostra coletada de acordo com os critérios da Sociolingüística laboviana (LABOV, 1972); e uma segunda, coletada em contextos interacionais livres, ou seja, sem o controle de qualquer variável, e de modo secreto (RONCARATI, 1996). Denominamos tecnicamente o primeiro tipo de *Amostra Censo* ou *Amostra Comunidade* (AC, daqui em diante), e o segundo tipo, de *Amostra de Interação* (AI, daqui em diante).

Em princípio, esses dois tipos de amostras servem a propósitos diferenciados. AC constitui um tipo mais propício para estudos sociolingüísticos e outros que necessitem de comprovação empírica, e AI, para estudos mais voltados para a interface gramática/discurso, uma vez que sob tal abordagem, concebe-se que a codificação lingüística do falante é uma decisão que decorre de um modelo de interação verbal construído na interlocução, ou mais precisamente no discurso (HENGEVELD, 2002).

No que se refere aos aspectos éticos da pesquisa, todos os informantes que forneceram dados lingüísticos para esse projeto expressaram, de antemão, sua concordância em participar da pesquisa, cujos objetivos lhes foram claramente explicitados.

---

[5] Um banco de dados anotado implica disponibilizar os arquivos sonoros acompanhados de suas respectivas transcrições ortográficas e outras notações pertinentes para o entendimento dos arquivos sonoros.

[6] Devemos fazer menção aqui à base de dados, com sede na UNESP/Araraquara, que registra a língua escrita contemporânea do Brasil, a partir de gêneros discursivos variados. Esse banco de dados já serviu a importantes obras de referência na lingüística brasileira (v. NEVES, 2000; BORBA, 2002).

[7] O nome *IBORUNA* (= Rio Preto) tem motivação histórica; é um topônimo de origem tupi-guarani que se pretendeu atribuir a cidade de São José do Rio Preto por ocasião da comemoração do seu cinqüentenário. A contundente intervenção do episcopado riopretense não só impediu a mudança como conquistou de maneira definitiva a denominação primitiva, São José do Rio Preto, reduzida a Rio Preto de 1906 a 1944.

3.1. A Amostra Censo ou Amostra Comunidade

Na constituição da AC, elegemos as variáveis sociais comprovadamente relevantes para os estudos sociolingüísticos (v. MOLLICA, 2003). Assim, para garantir a representatividade do censo lingüístico, controlamos as variáveis elencadas no quadro 1, dado a seguir.

Do cruzamento das variantes resultou um total de 160 células, que definiram o perfil social dos informantes. Excluídas oito células impossíveis de serem preenchidas (faixa etária de *7 a 15 anos X* escolaridade *Ensino Superior*), definem-se, então, os 152 informantes.

Quadro 1: Variáveis controladas na constituição da Amostra Censo.

Variáveis [8]	Variantes
1. Sexo	(1) masculino; (2) feminino
2. faixa etária	(1) de 7 a 15 anos; (2) de 16 a 25 anos; (3) de 26 a 35 anos; (4) de 36 a 55 anos; (5) + 55 anos
3. escolaridade	(1) 1° Ciclo do Ens. Fundamental; (2) 2° Ciclo do Ens. Fundamental; (3) Ens. Médio; (4) Ens. Superior
4. Renda	(1) + 25 sal. mínimos; (2) 11 a 24 sal. mínimos; (3) 6 a 10 sal. mínimos; (4) até 5 sal. mínimos.

A inclusão de, no mínimo, cinco informantes por célula elevaria esse número para 760 informantes, o que dificultaria a exeqüibilidade do projeto, principalmente pelo seu custo. A manutenção do número de 152 informantes não compromete a representatividade da amostra, como já bem demonstraram outros projetos e o próprio Labov (1972), ao considerar que a variação é bastante padronizada e, mesmo não havendo um imenso número de falantes para sua comprovação, a regularidade lingüística emerge, o que autoriza generalizações acerca da língua usada na comunidade como um todo. Entretanto, como adverte Paiva (1999, p. 7), essas implicações, "embora não possam ser ignoradas, não chegam a comprometer o estudo sociolingüístico" 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.

A inclusão de *área geográfica* como mais uma variável de estratificação faria crescer consideravelmente o número de informantes. Assim, para a execução desse censo lingüístico, os 152 informantes foram distribuídos proporcionalmente por entre as sete cidades da região, considerando a densidade populacional mostrada no quadro 3 a seguir.

Quadro 2: Distribuição dos informantes proporcionalmente à densidade populacional das cidades da região.

Cidades da Região de São José do Rio Preto	População*	Número de informantes
1. Bady Bassit (12 km ao sul de SJRP)	11.475	04
2. Cedral (14 km, ao sul de SJRP)	6.690	02
3. Guapiaçu (16 km , ao leste de SJRP)	14.049	05
4. Ipiguá (18 km, ao norte de SJRP)	3.461	01
5. Mirassol (14 km, a oeste de SJRP)	48.233	16
6. Onda Verde (25 km, ao norte de SJRP)	5.407	02
7. São José do Rio Preto[9]	357.705	122
Total da população representada	447.020	152

* Fonte: IBGE (Censo 2000)

A definição do perfil social desses informantes por cidade seguiu o método da distribuição aleatória (SILVA, 1996, 2003) por área geográfica, que consiste nos seguintes procedimentos: (i) em uma primeira urna, colocamos todos os perfis sociais, identificados pelos números de 1 a 152 (v. quadro 4 abaixo); (ii) em uma segunda urna, colocamos os nomes de todas as cidades da região; (iii) por meio de escolha ao acaso, retirávamos de cada uma das urnas um perfil social e um nome de cidade, definindo assim a origem geográfica do perfil social; (iv) em seguida, voltávamos apenas para segunda urna o nome da cidade escolhida, até que se atingisse o seu número de informantes; (v) nova escolha combinada era feita, até ser definida a origem geográfica de todos os informantes. Esse procedimento de escolha aleatória garantiu-nos a probabilidade de que quaisquer dos perfis sociais pertencessem a uma dada cidade. Esses resultados podem ser mais bem visualizados no quadro 4 apresentado na página seguinte.

[8] Cabe aqui uma rápida justificativa acerca da proposição de algumas variantes: no recorte contínuo de *faixas etárias*, a primeira representa a fase em que os padrões lingüísticos estão ainda em fixação; da segunda em diante, a motivação é buscada na pressão social sobre a linguagem do indivíduo, variável de acordo com a inserção no mercado de trabalho (cf. SILVA, 1996); a segmentação das variantes de *escolaridade*, embora pedagogicamente extinta nos dois primeiros ciclos, preserva diferenças salientes em termos de currículo e metodologia de ensino e representa a divisão escolar em vigor no tempo em que a maioria dos informantes se enquadra; sobre a variável *renda familiar*, reconhecendo a dificuldade de divisão de classe social por indicadores diversos, optamos apenas por esse indicador.

[9] Inclui os Distritos de Talhados e Engenheiro Schmidt .

Definidos os perfis sociais, a preparação de AC, incluindo a localização de um determinado perfil social, a gravação da entrevista e a posterior transcrição, ficou exclusivamente sob a responsabilidade de 17 integrantes da equipe técnica, aos quais coube o contato com uma média de 10 informantes. A tarefa de coleta das entrevistas foi precedida por uma fase piloto de intenso treinamento da equipe.

Seguindo roteiro previamente elaborado, as entrevistas da AC do Projeto ALIP foram direcionadas para obtenção de cinco tipos de textos de cada informante, com base na metodologia exposta em Votre & Oliveira (1995), a saber: *narrativa de experiência pessoal*, *narrativa recontada*, *relato de descrição*, *relato de opinião* e *relato de procedimento*. A razão para a obtenção desses "gêneros" textuais deve-se, em grande parte, à dificuldade de suas delimitações em textos orais, quando se deseja correlacionar determinadas manifestações lingüísticas a esses gêneros. Seria ingenuidade de nossa parte, entretanto, esperar a obtenção de textos genuinamente narrativos, opinativos, injuntivos e descritivos (v. TRAVAGLIA, 2002). Assim, por meio de roteiro de entrevista previamente definido, acreditamos na **predominância** dos gêneros textuais pretendidos.

Quadro 3: Distribuição dos perfis sociais dos informantes da Amostra Censo por cidades da Região

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

Legenda: Bady Bassit; Cedral; Guapiaçu; Onda Verde; Mirassol; Ipiguá; São José do Rio Preto

Somente para efeito de comparação em relação ao tamanho da amostragem de AC, listamos, no quadro 5, alguns bancos de dados provenientes de outros projetos. Desse quadro, mais do que o tempo de constituição das amostras, interessa destacar, na comparação com o banco de dados do Projeto ALIP, a quantidade de informantes envolvidos para o tamanho e a densidade demográfica das respectivas regiões em que o censo se realizou.

Quadro 4: Quadro comparativo de amostras de fala de diferentes projetos: local, tamanho, variáveis, tempo de constituição.

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.
Discurso & Gramática (CUNHA, 1998)	Natal	20	Sexo, faixa etária e escolaridade	Não há referência a prazos.
Dialetos Sociais Cearenses (ARAGÃO e SOARES, 1996)	Fortaleza	23	Bairro, sexo, faixa etária, escolaridade e classe social	Não há referência a prazos.
PEUL – Amostra Comunidade (SCHERRE, 1996)	Cidade do Rio de Janeiro	64 (48 + 16 ampliados)	Sexo, faixa etária, escolaridade	Iniciado em 1980 e concluído em 1983.
LUAL (MOURA, 1995)	Estado de Alagoas	32 horas	Sexo, faixa etária, escolaridade	1991 (até 1995 não estava concluído)
VARSUL (VANDRESSEN, 1995)	Região Sul (12 áreas)	288 (24 por área)	Sexo, faixa etária e escolaridade	1991 (até 1995, não estava concluído)

Resumidamente, a preparação definitiva das amostras de AC compreendeu os seguintes passos:

(i) **Treinamento dos entrevistadores**, que envolveu a exposição de questões teórico-metodológicas e questões práticas, encaminhadas pelos coordenadores e/ou professores convidados, referentes à execução do censo lingüístico.
(ii) **Fase piloto de coleta**, na qual, cada um dos entrevistadores coletou uma entrevista, que teve avaliada sua adequação (qualidade sonora, quantidade de material, tipo de gênero solicitado, quantidade de intervenção do entrevistador[10] etc), momento também em que aprendeu a lidar com gravador digital e com o armazenamento dos arquivos de som eletronicamente.
(iii) **Elaboração de roteiro prévio da entrevista**, de acordo com o perfil sócio-cultural do informante. Preferencialmente era feito contato prévio com o informante antes da elaboração do roteiro da entrevista para a coleta da amostra de fala.
(iv) **Coleta definitiva**, para a qual a escolha dos informantes ficou a critério do entrevistador, observado o perfil social. Cada informante produziu cinco tipos de texto oral, de acordo com os gêneros discursivos definidos, seguindo-se, aproximadamente o seguinte roteiro: (i) *narrativa de experiência pessoal*, envolvendo relatos de história alegre ou triste; (ii) *narrativa recontada*, envolvendo história alegre ou triste ocorrida com outrem, evitando-se relato de filmes, novelas etc; (iii) *texto descritivo*, envolvendo descrição de local; (iv) *relato de procedimentos*, baseado em experiências culinárias ou outra atividade que exija procedimentos ordenados; (v) *relato de opinião*, abordando temáticas variadas, dependendo da faixa etária do informante (escola, família, religião, relacionamentos afetivos, esportes, problemas sociais etc). A coleta tentou dirimir a artificialidade da situação comunicativa, atentando-se para a condução da entrevista de uma forma mais natural e descontraída.
(v) **Elaboração do relatório de coleta**, fase durante a qual o entrevistador elaborou *Diário de Campo* e *Ficha Social* com dados do informante.
(vi) **Armazenamento eletrônico e transcrição das amostras**: todas os arquivos sonoros armazenados nos gravadores digitais foram transferidos para os computadores da Sala do Projeto ALIP, para posterior transcrição.

3.2. A Amostra de interação

O Projeto ALIP tinha como proposição inicial o registro de 10 amostras de interação social, gravadas a partir da interação dialógica entre 10 diferentes pares de interlocutores, cujos perfis sociais seriam estabelecidos por diferentes graus de assimetria social, conforme mostrado no quadro 6, a seguir.

[10] O entrevistador, na coleta de AC, deve ser apenas um estimulador de maior quantidade de material lingüístico e não um interlocutor ativo.

Quadro 5: Perfis sociais inicialmente propostos para a constituição de AI

Graus de assimetria	Característica social do par em interação	
Zero	Homem, 26 a 35 anos, 1° Ciclo, Classe A	Homem, 26 a 35 anos, 1° Ciclo, Classe A
	Mulher, 26 a 35 anos, 1° Ciclo, Classe A	Mulher, 26 a 35 anos, 1° Ciclo, Classe A
1	Homem, 26 a 35 anos, 1° Ciclo, Classe A	Mulher, 26 a 35 anos, 1° Ciclo, Classe A
	Homem, 26 a 35 anos, 1° Ciclo, Classe D	Mulher, 26 a 35 anos, 1° Ciclo, Classe D
2	Homem, 26 a 35 anos, 1° Ciclo, Classe A	Mulher, + 55 anos , 1° Ciclo, Classe A
	Homem, + 55 anos , 1° Ciclo, Classe D	Mulher, 26 a 35 anos, 1° Ciclo, Classe D
3	Homem, 26 a 35 anos, 1° Ciclo, Classe A	Mulher, + 55 anos , Superior, Classe A
	Homem, + 55 anos, Superior, Classe D	Mulher, 26 a 35 anos, 1° Ciclo, Classe D
4	Homem, 26 a 35 anos, 1° Ciclo, Classe A	Mulher, + 55 anos , Superior, Classe D
	Homem, + 55 anos, Superior, Classe D	Mulher, 26 a 35 anos, 1° Ciclo, Classe A

Essa proposta, entretanto, mostrou-se inviável, em vista dos objetivos esperados com a obtenção dessas amostras, qual seja: a espontaneidade de conversações mais próximas da fala coloquial e distensa, aquelas que o falante produz em seu dia-a-dia. Além disso, a pouca disposição dos informantes inicialmente contactados em contribuir com a constituição desse tipo de amostra foi também um fator que exigiu uma redefinição da metodologia para a sua obtenção. Assim, RONCARATI (em comunicação pessoal), pioneira em montar um banco de dados interacional (RONCARATI, 1996), sugeriu-nos que as gravações fossem feitas secretamente, atentando-se para situações de interação diversas, única maneira de se garantir a naturalidade do contexto interacional. E assim a equipe técnica passou a proceder, dando preferência para contextos mais livres, como conversas em contextos familiares, e atentando-se para a escolha de uma situação em que a qualidade sonora da gravação pudesse ser mantida, situação que exigiu dos documentadores a posse constante do gravador. Diante dessa nova estratégia, os perfis sociais para a composição de AI passaram a ser livres, sem o controle de qualquer variável, incluindo aí o número de participante da interação dialógica.

Como forma de manutenção dos aspectos éticos da pesquisa, preocupação que guiou toda a coleta de AC, ao final da gravação de uma AI, todos os envolvidos na interação eram cientificados da coleta e dos seus objetivos, bem como da estratégia adotada, sendo facultado a eles o direito de não ceder a sua amostra de fala. Em havendo concordância, era então coletadas a ficha social dos informantes e a assinatura da declaração livre e consentida para a cessão de suas amostras de fala ao Projeto ALIP. Em havendo recusa, a orientação era a de que os arquivos gravados fossem apagados imediatamente na frente do informante. Das gravações realizadas, não houve rejeição por parte de nenhum dos informantes gravados secretamente, uma vez que o Projeto ALIP já se tornara bastante conhecido na comunidade, dada a sua repercussão na mídia impressa e televisiva.

Pautadas por essas orientações metodológicas, foram coletadas 10 amostras de interação, cujas descrições são dadas no quadro 7.

Quadro 6: Informações sobre as interações dialógicas de AI.

AMOSTRA	CONTEXTO DE INTERAÇÃO – INFORMANTES ENVOLVIDOS
AI-001-CAS	Conversa em contexto familiar, envolvendo cinco informantes de diferentes perfis sociais.
AI-002-GIL	Conversa entre duas amigas vizinhas, de perfil social semelhante, em frente a casa de uma delas.
AI-003-ILHA	Conversa entre tia e sobrinha, de diferentes faixas etárias, em contexto familiar
AI-004-OND	Conversa entre duas irmãs de níveis de escolaridade diferentes.
AI-005-CAS	Conversa informal, em ambiente universitário, entre duas informantes de mesmo perfil social.
AI-006-MAR	Conversa em contexto familiar entre quatro informantes de diferentes faixas etárias
AI-007-FER	Conversa em contexto familiar entre dois informantes de sexo/gêneros diferentes.
AI-008-CAM	Conversa entre amigos, de sexo/gêneros diferentes, em ambiente universitário.
AI-009-CAS	Conversa entre advogado e cliente, de perfis sociais diferentes, em escritório.
AI-010-CAS	Discussão entre dois advogados, de mesmo perfil social, sobre uma peça jurídica

A preparação definitiva das gravações de AC envolveu: (i) a coleta de amostras de fala em contexto livre de interação; (ii) a elaboração do relatório de coleta; (iii) o armazenamento e a transcrição da amostra coletada.

3.3. Da validação do material coletado

Todas as gravações de AC e de AI, até o ponto definitivo da constituição do banco de dados, foram submetidas a três fases de validação, conforme seguem abaixo descritas.

(i) Primeira validação: anteriormente à transcrição dos dados, todas as gravações foram ouvidas para avaliação da adequação do material coletado (qualidade das gravações, naturalidade das entrevistas e diálogos, quantidade de intervenções do entrevistador, fidelidade aos tipos de gêneros discursivos etc). Gravações invalidadas foram coletadas novamente, com substituição de informantes, se necessário.

(ii) Segunda validação: envolveu a checagem do material transcrito no cotejo com as respectivas gravações. Essa tarefa foi executa por meio da troca das transcrições entre os transcritores. Aleatoriamente, os coordenadores do

projeto também checaram os materiais validados nessa fase. A não correspondência entre gravação e transcrição implicou a refacção do trabalho.

(iii) Validação Final: envolveu a checagem da fidelidade do material transcrito no confronto com as gravações. Para garantir a homogeneidade das amostras definitivas, essa tarefa ficou restrita a uma equipe de três membros do grupo, com uma última validação de todas as transcrições por um quarto membro.

3.4. Da transcrição das amostras de fala

Ao se elaborar um sistema de transcrição da língua falada, é importante ter claro que o seu objetivo básico "é transpor o discurso falado, da forma mais fiel possível, para registros gráficos mais permanentes" (PAIVA, 2003, p.135). Mas essa fidelidade é 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 influenciam a percepção dos dados lingüísticos (cf. EDWARDS & LAMPERT, 1992). Portanto, a elaboração de um sistema de transcrição implica tomada de decisões teoricamente embasadas e que devem ser explicitadas. Esse foi o nosso objetivo, ao elaborarmos, conjuntamente com a equipe de transcritores, o manual de transcrição do Projeto ALIP, ou seja, deixamos claros os princípios organizadores do sistema de transcrição.

Nosso sistema de transcrição foi constituído a partir da comparação das convenções usadas por outros projetos semelhantes. Procuramos, no entanto, aprimorar essas convenções por meio da categorização das normas que as regulamentam, com o objetivo de explicitar a natureza do que se está sendo transcrito. As categorias mais gerais por nós adotadas foram: (i) convenções sobre a grafia das palavras; (ii) convenções sobre alguns elementos prosódicos; (iii) convenções sobre a indicação de alguns aspectos da interação; (iv) convenções sobre os comentários do transcritor. O quadro 8, à página seguinte, apresenta os sinais usados para a transcrição. Discutimos, a seguir, as categorias propostas e o sistema de transcrição adotado.

As convenções sobre a grafia das palavras que encontramos são, relativamente, as mais freqüentes nos sistemas de transcrição propostos para o Português Brasileiro. Talvez isso se deva ao fato de essas convenções serem baseadas principalmente nas convenções ortográficas usadas para a língua escrita. No entanto, alguns aspectos são anotados de modo diferente do que é prescrito pelas regras gramaticais, como por exemplo, o emprego de iniciais maiúsculas restrito a nomes próprios. Ainda sob essa categoria, estão arrolados truncamentos[11] e metalinguagem do informante por contemplarem, em certa medida, aspectos da grafia do que é transcrito da língua oral.

[11] *Truncamento* indica a ocorrência de palavras incompletas ou interrupção brusca.

Quadro 7: Categorias gerais do manual de transcrição das amostras de fala

OCORRÊNCIAS	SINAIS	Observações / Exemplos
Sobre a grafia das palavras		
Nomes próprios	Iniciais Maiúsculas	...a festa foi na casa do ***Carlos***?...
Nomes de obras (livros, revistas, jornais) ou palavras estrangeiras	Grafia da língua de origem e em itálico	... adorava ouvir ***Purple Rain***..
Marcadores discursivos	Ocorrência seguida do ponto "?"	... é pra deixar aqui **né?...**
Fáticos/Interjeições	Ocorrência seguida do ponto "!"	... **ah!** ... que alívio...
Numerais	Por extenso	... foram **trinta e três** alunos...
Truncamento	/	... **ca/** casou semana passada...
Metalinguagem do informante	'aspas simples'	... o **'mesmo'** do carioca
Citação	"aspas duplas"	... Armstrong disse **"pequeno passo para o homem... gigantesco salto para a humanidade"**
Sobre elementos prosódicos		
Silabação	Hífen entre as sílabas (sem espaço)	... foi quando ele disse... – "**fi-que-a-qui**" – -...
Pausa (longa ou breve)	Reticências	**...** ele**...** voltou feliz**...**
Ênfase	CAIXA ALTA	... ele almoçou com **ELA**...
Alongamento (vogais/consoantes)	Dois pontos seguidos vezes	... ele **a::**cha...
Pergunta	Ponto-de-interrogação	... você vai à festa**?**...
Sobre alguns aspectos da interação		
Identificação dos participantes da interação	Documentador (Doc.) Informante (Inf) e Interveniente (Int)	
Início de turno	Minúsculas	Doc.: **O** senhor poderia me contar sobre .... Inf.: **F**az uns dois meses.... Int.: **N**ão ... foi no mês passado
Discurso direto	Aspas duplas e duplo travessão	... ela disse **-"Vamos à festa?"**- eu respondi **-"talvez"**-
Mudança do fluxo discursivo	Duplo travessão	... eu não tinha **– fique quieto!** ((falando com o cachorro)) **–** tempo de estudar...
Superposição/simultaneidade de vozes	Texto entre colchetes, com índice sobrescrito à esquerda do colchete inicial[12]	Inf.1: eu não tinha saído de lá... **[1][e foi então...]** Doc.: **[1][cê tava]** em casa ainda?
Sobre os comentários do transcritor		
Hipótese do que se ouviu	Entre parênteses	... aí ele **(fez)** ...
Comentário descritivo do transcritor	Entre parênteses duplos	**((risos)) ((tossiu)) ((ruído)) ((ininteligível))**

(in: Gonçalves & Tenani, 2004)

Um outro conjunto de convenções trata principalmente de elementos suprassegmentais, particularmente alguns elementos prosódicos,[13] como pausa, duração (alongamento de vogais e/ou consoantes), ritmo e velocidade de fala (silabação),[14] entoação (somente o padrão da interrogativa direta) e variação de altura e intensidade percebida como 'ênfase'.[15] Sobre cada um desses elementos prosódicos, várias observações poderiam ser feitas, mas nos restringimos a 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 uma certa facilidade de percepção (auditiva, em geral). 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 versa sobre o grau de detalhamento dos elementos prosódicos transcritos. Ainda sobre a pausa, observa-se a opção por somente indicar a ocorrência, sem se ocupar com a duração de cada uma das ocorrências.

A partir dessas reflexões, podemos explicitar o desafio que enfrentamos quanto a esse aspecto da transcrição nos seguintes termos: é suficientemente adequado propor que a seleção dos elementos prosódicos e que o grau de detalhamento da transcrição desses elementos selecionados sejam baseados na (relativa facilidade de) percepção do transcritor (falante nativo da variedade do português analisado)? A solução adotada inicialmente foi transcrever os elementos prosódicos mais freqüentemente anotados (nos sistemas de transcrição considerados) e com o mesmo grau de refinamento. Um cuidado extra consistiu na realização de

[12] A função do índice é a delimitar a parte de turnos de fala sobrepostos.

[13] Os elementos prosódicos desempenham diferentes funções lingüísticas (v. CAGLIARI, 1992).

[14] Ritmo e velocidade de fala são elementos prosódicos distintos, embora muito freqüentemente confundidos.A rigor, a velocidade de fala 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 *allegro*). (v. MORAES & LEITE, 1993).

[15] Nos sistemas de transcrição consultados, a ênfase sempre é indicada, mas nunca explicitamente. Neste texto não teceremos considerações sobre a ênfase.

um treinamento a fim de explicitar técnicas de transcrição de base auditiva somadas a recursos de programas de análise de fala, como o PRAAT, de maneira que houvesse homogeneização do material transcrito.

A terceira categoria proposta diz respeito à indicação de *aspectos da interação*: identificação do turno da conversação (documentador e informante), intervenções ocasionais e sobreposição de vozes. Para assegurar a identificação dos participantes da interação, optamos também por indicar 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 seqüência temática*, que requer uma análise prévia do texto oral.[16]

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[17] a serem expressos).

4. Das dificuldades encontradas na execução do Projeto

De importância central para este artigo, é o relato das dificuldades encontradas durante a constituição do banco de dados Iboruna e as soluções para cada uma delas, possíveis de serem alcançadas somente com o desenvolvimento do projeto. O certo é que esforços foram envidados para a manutenção da proposta original. Passemos a discorrer brevemente sobre cada uma delas.

4.1. O uso dos gravadores digitais

Na constituição do banco de dados, um primeiro problema de difícil solução e que demandou tempo para sua resolução foi encontrar a configuração adequada dos microcomputadores para a manipulação dos *softwares* de transferência dos arquivos de som dos gravadores para os computadores. Os manuais que acompanham os gravadores digitais nem sempre trazem instruções claras sobre a configuração da "porta" das máquinas (Porta COM) que permitem acessar os arquivos armazenados nos gravadores.

Esse problema seria facilmente contornável pelo recurso a gravadores digitais com mini-disco, em que a gravação é armazenada diretamente em mini-discos, os quais podem, posteriormente, ser copiados diretamente para os microcomputadores. Esse recurso, entretanto, não foi empregado, em vista do alto custo desses gravadores, em média quatro vezes mais que os gravadores que armazenam gravações em memória. Além disso, deve-se prever ainda o alto custo com a aquisição de mini-discos, dado o número de informantes que integram a amostra do projeto.

Some-se a esse problema, a constante necessidade de apagar dos gravadores as entrevistas gravadas, após a sua transferência definitiva para as máquinas. Esta é uma medida que adotamos, em vista do risco que se corre em deixar uma gravação armazenada somente na memória do gravador até que se complete todas as oito horas de gravação permitidas.[18]

Mesmo com alguns cuidados, outros problemas surgiram com esses equipamentos, sendo o principal deles a presença constante de ruídos, após a transferência do arquivo de som para as máquinas do projeto. A audição da gravação diretamente do gravador não apresentava problemas, mas após a conversão do arquivo armazenado na memória do gravador para o formato .WAV (*wave*), surgiam "chiados", sons "raspados", trechos inaudíveis, em longas partes das gravações. Esse foi um dos problemas que levou à invalidação de várias entrevistas, mesmo de qualidade textual boa. O problema foi solucionado com a aquisição de novos gravadores, do tipo *MP3 Player*, que produz arquivos de voz diretamente no formato .WAV, sem a necessidade de conversão. O dispositivo no qual se armazenam os arquivos funciona como um *pen-drive*, que, inserido em uma porta *USB*, permite a transferência (=cópia) dos arquivos diretamente para a máquina. Essa solução, embora tardia, permitiu a continuidade das coletas sem o comprometimento posterior das gravações.

Apesar desses problemas, foi possível levar a bom termo a obtenção dos arquivos sonoros definitivos. A qualidade sonora das gravações foi mantida, como era nossa preocupação, ao contestarmos as gravações analógicas. Sem sombras de dúvidas, obtivemos uma boa qualidade sonora, superiormente melhor àquelas obtidas analogicamente, mas ainda não pudemos assegurar uma qualidade ideal, a exemplo de gravações

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

[17] Comentário 'ininteligível' é usado quando não for possível assegurar o que foi enunciado.

[18] Os gravadores utilizados são das marcas *gama-power* e *power-pack*.

realizadas em laboratórios experimentais, tendo em vista a metodologia que o projeto emprega para a realização do censo lingüístico.

4.2. A etapa da transcrição das gravações

A meta esperada era a de que cada integrante da equipe técnica produzisse a transcrição completa das gravações por que ficou responsável. Entretanto, essa meta foi alcançada parcialmente, em razão do esforço de concentração e do elevado tempo que essa árdua tarefa demanda.

Além disso, o treinamento dos alunos para a transcrição tomou um tempo considerável, o que ocasionou certo atraso na execução do projeto. Nem sempre os transcritores apresentaram total segurança na aplicação das convenções estabelecidas para a transcrição, principalmente quando se deparavam com situações não previstas no manual.

Como somente ao final do projeto pudemos medir a real dimensão desse trabalho, elegemos, em etapa posterior uma equipe menor, com tarefa exclusiva de revisão das transcrições feitas, de modo a garantir uma homogeneidade mínima na aplicação das regras contidas no manual de transcrição.

4.3. A atuação da equipe técnica

Um outro problema detectado durante a constituição do banco de dados diz respeito à uniformidade dos procedimentos metodológicos da equipe técnica, desde a obtenção das amostras de fala até a transcrição das gravações.

Parece-nos, hoje, ingenuidade pensar em homogeneidade de procedimentos, quando se lida com uma equipe grande de pessoas (SILVA, 2003). A garantia dessa homogeneidade só seria alcançada se toda tarefa de montagem do banco de dados estivesse a cargo de um único pesquisador, o que é praticamente impossível, dada a dimensão de um censo lingüístico.

Apesar do treinamento e da supervisão a que a equipe técnica se submeteu, constantes desvios de procedimentos foram detectados, alguns com certa insistência, fato que também retardou o andamento do projeto. Cite-se, como principal problema, a invalidação de gravações resultantes de "gêneros" textuais inadequados, por vezes atribuída à desatenção do documentador na condução das entrevistas.

Uma primeira medida para contornar esse problema foi a elaboração de manuais de todas as etapas do projeto. Uma outra solução foi a substituição de alguns membros da equipe técnica e/ou a ampliação da equipe, medida que se esbarrou no oferecimento de novo treinamento. Essa, entretanto, nos pareceu a solução mais viável, com vistas à garantia de um banco de dados de qualidade.

Devemos deixar enfatizada a importância da integração de todos os envolvidos com a elaboração de um banco de dados, sem a qual, certamente, não teríamos atingido os resultados hoje apresentados.

4.4. A localização de perfis sociais da Amostra Censo

A última e mais problemática dificuldade com a qual nos deparamos refere-se à localização de informantes com perfis sociais determinados, uma vez que algumas variáveis sociais se co-determinam. Sobre essa dificuldade, algumas medidas foram adotadas em comum acordo com a equipe técnica.

Mesmo com toda a divulgação do projeto na mídia impressa e televisiva, alguns perfis sociais continuavam sendo de difícil localização, dados os critérios impostos pela metodologia sociolingüística. Uma solução para tornar viável a localização de todos os 152 perfis sociais foi passarmos a desprezar a variável *renda familiar*, pois esse era o principal critério social que, no intercruzamento com a variável *escolaridade*, dificultava a localização dos informantes na comunidade. Por exemplo, tornava-se difícil identificar informantes com baixa escolaridade (1^{o}. ou 2^{o}. Ciclo do Ensino Fundamental) e com alta renda familiar (mais de 25 salários mínimos). Do mesmo modo, de também difícil localização eram informantes de nível superior com baixa renda familiar (até cinco salários mínimos), embora este perfil fosse mais provável de ser encontrado. Nas reuniões de treinamento com a equipe técnica, quando já contávamos com pouco mais de 70% dos informantes já gravados, decidimos desconsiderar a variável *renda familiar* se esse fosse realmente o impedimento para as gravações que ainda restavam, decisão que acelerou, nos últimos meses, o trabalho da equipe técnica.

Essa solução não invalidou em nada o trabalho até então realizado, uma vez que, no rigor sociolingüístico até então prezado, passamos a contar com quatro informantes por célula social, o que faz manter o número de 152 informantes, como previsto inicialmente. Entretanto, mesmo tendo sido desconsiderada essa variável, um outro entrecruzamento, também determinante das relações sociais e que

vinha dificultando o cumprimento do projeto na íntegra, era devido à variável *escolaridade*, que, no cruzamento com a variável *faixa etária*, tornava bastante difícil a localização de informantes das faixas etárias intermediárias (16 a 25 anos e de 26 a 35 anos) com baixo nível de escolaridade (1º. Ciclo do Ensino Fundamental), últimos perfis sociais gravados.

Some-se a essas dificuldades o fato de os informantes terem de ser nativo das cidades onde o censo lingüístico se realizou ou de nela estar vivendo desde pelo menos os cinco anos de idade. Por ser um problema de mais difícil solução, uma vez que pode ser reflexo da própria constituição social da comunidade,[19] apenas relativizamos esses critérios, da seguinte forma: (i) no perfil de informantes do 1º. Ciclo do Ensino Fundamental (1ª. a 4ª. série) e do 2º. Ciclo do Ensino Fundamental (5ª. a 8ª. série) poderiam estar incluídos, respectivamente, informantes que estivessem também cursando a 5ª. série e o 1ª. série do Ensino Médio, ou nelas terem se estacionado, uma vez que a influência da escola nesses limites de níveis não seria tão grande assim, a ponto de se verificarem mudanças qualitativas de uma série para outra; (ii) se o critério "ser natural da cidade" ou "nela residir desde pelo menos os cinco anos" fosse impedimento (como foi em muitos casos de descarte de informante potencial) para a seleção do informante, essa idade poderia se estender para até os 10 anos, fase em que os padrões lingüísticos da comunidade de origem do informante podem ainda ser alterados em favor dos da nova comunidade (v. NARO, 2003).

Uma vez que o afrouxamento dos critérios sociolingüísticos só foi adotado como último recurso, quando o trabalho da equipe já se mostrava improdutivo, poucos informantes da AC não atenderam completamente o perfil social previamente determinado.

Outros problemas de ordem teórica e metodológica surgidos durante o desenvolvimento do projeto encontram-se registrados em Gonçalves & Tenani (2004, 2005).

5. Considerações finais

Embora não trabalhando dentro de todas as condições desejáveis, enfrentando problemas e dificuldades imprevisíveis, todos os objetivos iniciais do projeto ALIP foram plenamente alcançados. Em vista das motivações e dos objetivos que nortearam sua proposição, o banco de dados Iboruna é hoje uma realidade, comportando amostras de fala que servem a propósitos investigativos diversificados e que vêm instigando a pesquisa lingüística realizada no interior do estado de São Paulo.

Diante do propósito deste artigo, expressamos, nessas considerações finais, nosso desejo de partilhar nossa experiência e o material resultante da implantação do Projeto ALIP, que hoje segue com pesquisas de descrição do PB falado na sua variedade riopretana. Todas as amostras de fala estão disponíveis em meio eletrônico, na sede do Projeto ALIP, na UNESP de São José do Rio Preto. É nossa intenção colocar à disposição, gratuitamente, em *homepage* do projeto, todo o material integrante do banco de dados Iboruna, inclusive o material sonoro, em vista do investimento de recursos públicos na sua constituição.

6. Referências Bibliográficas

ARAGÃO, M.S.S., SOARES, M.E. (orgs.) *A linguagem falada em Fortaleza: diálogos entre informantes e documentadores (materiais para estudo)*.Fortaleza: UFC, 1996.

BORBA, F.S. *Dicionário de usos do português*. São Paulo: Contexto, 2002.

CAGLIARI, L. C. Prosódia: algumas funções do supra-segmentos. *Cadernos de Estudos Lingüísticos 23*, 1992: 137-152.

CASTILHO, A.T. Português culto falado no Brasil: história do Projeto NURC/BR. In: PRETI, D., URBANO, H. (orgs.) *A linguagem falada culta na cidade de São Paulo*: estudos. vol. 4, São Paulo: FAPESP/TAQ, 1990.

CASTILHO, A.T., PRETI, D. (orgs.) *A linguagem falada culta na cidade de São Paulo*: diálogos entre dois informantes. vol. 2, São Paulo: FAPESP/TAQ, 1987.

[19] No tocante a esse problema, não previmos na composição da amostragem uma distribuição dos informantes proporcional aos números absolutos da distribuição da população da região por critérios sociais como escolaridade, sexo, renda e faixa etária. Isso significa dizer, por exemplo, que, se a região possui um alto índice de habitantes com baixa escolaridade, o número de informantes com esse perfil social teria de ser maior e não igual aos de alta escolaridade. Preferimos, no entanto, manter uma distribuição equilibrada dos perfis sociais na amostragem constituída.

_____. *A linguagem falada culta na cidade de São Paulo*: elocuções formais. vol. 1, São Paulo: FAPESP/TAQ, 1986.

CASTRO, L.P.G. *O estudo da variante retroflexa na comunidade de São José do Rio Preto*. 2002. 106f. Dissertação (Mestrado) em Estudos Lingüísticos – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto.

CUNHA, M.A.F.C. (org.). *´Corpus´ discurso & gramática*: a língua falada e escrita na cidade de Natal. Natal: EDUFRN, 1998.

DIK, S.C. *The theory of Functional Grammar*. Parte I: The structure of the clause. Dordrecht: Foris, 1989.

GRAMÁTICA DO PORTUGUÊS FALADO. Campinas: UNICAMP/FAPESP, 8 volumes.

EDWARDS, J. A., LAMPERT, M. D. (eds) *Talking data: Transcription and coding in discourse research.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1992: 3-31.

GONÇALVES, S.L.G., TENANI, L.E. Projeto ALiRP: constituição de um banco de dados para estudo do português falado na região de São José do Rio Preto. *Revista Mosaico*, São José do Rio Preto, v.3, n.2, 2004, p. 13-40.

GONÇALVES, S.L.G., TENANI, L.E. Aspectos metodológicos na constituição de *corpora* de fala espontânea. *Estudos Lingüísticos*, São Paulo, v.34, 2005, p. 462-487.

HENGEVELD, K. The architecture of a Functional Discourse Grammar. In: MACKENZIE, J. L.; GÓMEZ GONZÁLES, M. A. (eds.) *A new architecture for Functional Grammar*. Berlin, New York: Mouton de Gruyter, 2002.

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

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

LIRA, S.A. *Variação lingüística urbana na região sul do país: censo*. Projeto apresentado ao CNPq., 1987.

MOLLICA, M.C., BRAGA, M.L. (Orgs.) *Introdução à Sociolingüística*: o tratamento da variação. São Paulo: Contexto, 2003.

MOLLICA, M.C. A relevância das variáveis não lingüísticas. In: MOLLICA, M.C., BRAGA, M.L. (Orgs.) *Introdução à Sociolingüística*: o tratamento da variação. São Paulo: Contexto, 2003, p. 28-31

MORAES, J. A. e LEITE, Y. F. Ritmo e velocidade de fala na estratégia do discurso: uma proposta de trabalho. In: ILARI, R. (org.) *Gramática do Português Falado*, v. 2, 1993: 67-77.

MOURA, M.D. A língua usada em Alagoas. Alagoas: EDUFAL, 1995.

NEVES, M.H.M. *Gramática de usos do português*. São Paulo: UNESP, 2000.

NARO, A.J. O dinamismo das línguas. In: MOLLICA, M.C., BRAGA, M.L. (Orgs.) *Introdução à Sociolingüística*: o tratamento da variação. São Paulo: Contexto, 2003, p. 43-50.

OLIVEIRA, D.P., ISQUERDO, A.N. A nova dialetologia: investigações e resultados. In: RONCARATI, C., ABRAÇADO, J. (org.) *Português brasileiro*: contato lingüístico, heterogeneidade e história. Rio de Janeiro: 7Letras, 2003, p. 50-54.

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

PAIVA, M.C. Transcrição de dados lingüísticos. In: In: MOLLICA, M.C., BRAGA, M.L. (Orgs.) *Introdução à Sociolingüística*: o tratamento da variação. São Paulo: Contexto, 2003, p. 135-146.

PAIVA, M.C., DUARTE, M.E.L. (org.) *Mudança lingüística em tempo real*. Rio de Janeiro: Contracapa, 2003.

PRETI, D., URBANO, H. (orgs.). *A linguagem falada culta na cidade de São Paulo*: entrevistas. vol. 3, São Paulo: FAPESP/TAQ, 1986.

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

SCHERRE, M.M.P. Breve histórico do Programa de Estudos Sobre o Uso da Língua. In: SILVA, G.M.O., SCHERRE, M.M.P. (orgs.) *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. (orgs.) *Padrões sociolingüísticos*: análise de fenômenos variáveis do português falado na cidade do Rio de Janeiro. Rio de Janeiro: Tempo Brasileiro, 1996, p. 51-81.

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

TRAVAGLIA, L.C. Tipos, gêneros e subtipos textuais e o ensino de língua materna. In: BASTOS, N.B. (org.) *Língua Portuguesa*: uma visão em mosaico. São Paulo: IP-PUC-SP/EDUC, 2002, p. 201-214.

VANDRESSEN, P. O Projeto Varsul – Avaliação e perspectivas sobre pesquisas do português falado na Região Sul. *Anais do I Encontro Nacional sobre Língua Falada e Ensino*. 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.