Problemas teórico-metodológicos na elaboração de um sistema de transcrição de dados interacionais: o caso do projeto ALIP (Amostra Lingüística do Interior Paulista)

Sebastião Carlos Leite Gonçalves
Luciani Ester Tenani

Recebido 30, jun. 2008/ Aprovado 17, set. 2008

Resumo

Tratamos, neste artigo, dos procedimentos metodológicos que nortearam os trabalhos de transcrição das amostras de fala do Projeto ALIP (Amostra Lingüística do Interior Paulista). Justificamos as opções pelas convenções registradas no sistema de transcrição, apontamos alguns problemas nas tarefas executadas pela equipe responsável pela transcrição e as possíveis soluções e, por fim, apresentamos alguns fenômenos lingüísticos já estudados com base nas amostras de fala transcritas.

Palavras-chave*: Transcrição. Amostra de fala. Banco de dados.*

Introdução

Neste artigo, como o próprio título sugere, temos por objetivo principal trazer à discussão os problemas teóricos e metodológicos, quando da elaboração do manual de transcrição que norteou a documentação das amostras de fala da região de São José do Rio Preto, e apresentar algumas soluções encontradas pelo Projeto na tentativa de homogeneização de procedimentos para a transcrição das amostras, em vista da dimensão do banco de dados e do elevado número de membros da equipe responsável pela documentação dessas amostras.

Para facilitar referências posteriores, nos referiremos ao Projeto de constituição do banco de dados e aos trabalhos de caracterização do português riopretano dele decorrentes como *Projeto ALIP* (Amostra Lingüística do Interior Paulista), e ao banco de dados, em si, como *Banco de dados Iboruna*.[1]

Esse artigo está dividido em três partes: na primeira, caracterizamos brevemente o banco de dados *Iboruna*, para, na segunda, abordar as questões teóricas e metodológicas das transcrições de dados interacionais do projeto ALIP. Na terceira e última parte, apresentamos alguns fenômenos lingüísticos abordados no interior do Projeto ALIP e o papel das transcrições na tarefa de caracterização desses fenômenos.

1. Breve caracterização do Projeto ALIP e do Banco de dados *Iboruna*

O projeto ALIP foi uma iniciativa concebida no interior do Grupo de Pesquisa em Gramática Funcional, (GPGF) da UNESP de São José do Rio Preto, entre os anos de 2002 e 2003, em razão do interesse dos membros do grupo na "Descrição Funcional do Português Oral e Escrito" e na "Variação e Mudança Lingüística", linhas de pesquisa que têm como principal diretriz o enfoque da língua usada no seu contexto social.

O projeto, que se constituiu sob dos auspícios da Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP (Proc. 03/08058-6) de 2004 a 2007, permitiu a constituição de um banco de dados com amostras do português falado na região noroeste do Estado de SP, mais especificamente na região delimitada por São José do Rio Preto e seis cidades que lhe fazem fronteira (*confira Fig. 1*).

[1] 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, para diferenciá-la de duas outras cidades homônimas de outros estados. A contundente intervenção do episcopado riopretano 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.

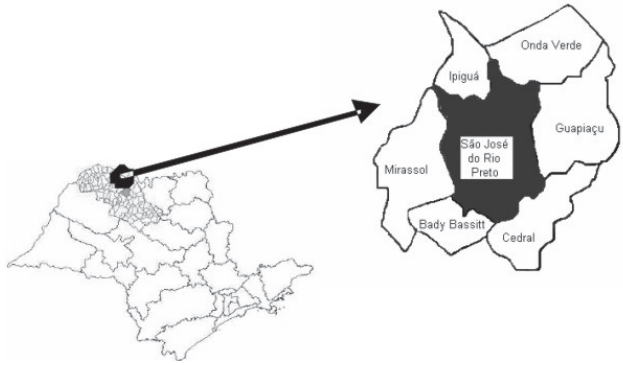

Fig. 1: Abrangência do Projeto ALIP
(Amostra Lingüística do Interior Paulista)

O banco de dados *Iboruna* foi idealizado para comportar dois tipos de amostras de fala:

(i) um primeiro tipo de amostra, tecnicamente denominado *Amostra Censo* ou *Amostra Comunidade* (AC) e coletado de acordo com os critérios da Sociolingüística laboviana (LABOV, 1972; VOTRE; OLIVEIRA; SILVA, 1995), envolveu o controle rigoroso das seguintes variáveis sociais:

a. sexo/gênero (masculino, feminino);

b. faixa etária, estratificada em 5 níveis (7 a 15 anos, 16 a 25 anos, 26 a 35 anos, 36 a 55 anos, + de 55 anos);

c. escolaridade, estratificada em 4 níveis (até 4 anos, de 5 a 8 anos, de 9 a 11 anos e + de 11 anos);[2]

d. renda familiar, estratificada em 4 níveis (até 5 salários mínimos, de 6 a 10 salários mínimos, de 11 a 25 salários mínimos, + 25 salários mínimos);

Do cruzamento das variantes de cada variável resultou um total de 152 células, que definiram os perfis sociais contactados na comunidade. Assim, AC compõe-se, portanto, de 152 amostras de fala, com duração aproximada de 40 minutos de gravação.

Seguindo roteiro previamente elaborado, as amostras de AC foram direcionadas para obtenção de cinco tipos de textos de cada informante, com base na metodologia exposta em Votre, Oliveira e Silva (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 delimitações de seqüências tipológicas em textos orais, quando se deseja correlacionar determinadas manifestações lingüísticas a gêneros textuais. Seria ingenuidade, entretanto, esperar a obtenção de textos

[2] Essas faixas de escolaridade correspondem, respectivamente, aos atuais 1º. Ciclo do Ens. Fundamental (até 4ª. série), 2º. Ciclo do Ens. Fundamental (até 8ª. série), Ens. Médio (até 3º. Colegial) e Ens. Superior.

genuinamente narrativos, opinativos, injuntivos e descritivos (cf. TRAVAGLIA, 2002). Assim, por meio de roteiro de entrevista previamente definido, espera-se a sempre a *predominância* dos tipos textuais pretendidos.

(ii) um segundo tipo de amostra, tecnicamente denominado *Amostra de Interação* (AI), coletado em contextos interacionais livres, sem o controle de qualquer variável, caracteriza-se pela gravação secreta (RONCARATI, 1996), preservando, assim, a total naturalidade da conversação. Para esse tipo de amostra foram gravadas 11 interações dialógicas, com uma média de 15 minutos de gravação cada uma, envolvendo de 2 até 5 informantes.

Em princípio, esses dois tipos de amostras servem a propósitos diferenciados. AC constitui um tipo mais propício a estudos da variação e da mudança lingüística em tempo aparente, e AI, a estudos mais voltados para a interface gramática/discurso, uma vez que sob tal abordagem, se concebe que a codificação lingüística do falante é uma decisão que decorre de um modelo de interação verbal naturalmente construído na interlocução, ou mais precisamente no discurso (HENGEVELD, 2002). Assim, enquanto o paradoxo do observador (LABOV, 1972) interfere, de certa forma, na naturalidade das interações de AC, ele é totalmente nulo em AI. Assim é que AC e AI compõem dois *corpora* de fala diferenciados, que, juntos, totalizam pouco mais de 1 milhão de palavras, provenientes de cerca de 100 horas de gravação de fala.

No que se refere aos aspectos éticos da pesquisa, todos os informantes cederam de livre e espontânea vontade suas amostras de fala, expressando, por meio de termo de declaração, sua adesão livre e consentida ao projeto, cujos objetivos lhes foram claramente explicitados.

Dado o espaço de que dispomos, não nos ateremos aos procedimentos metodológicos da constituição de cada uma das amostras. Passamos então a focar apenas questões de ordem teórica e metodológica que nortearam a tarefa de transcrição das amostras do Projeto ALIP, os obstáculos enfrentados na aplicação de um sistema de transcrição pela equipe técnica do Projeto e as soluções buscadas para minimizar esses obstáculos.

Sistema de transcrição e homogeneização de procedimentos

Coletadas as amostras de fala, trabalho bastante árduo é a transcrição ortográfica das gravações, que exige, além de aparatos tecnológicos adequados, documentadores com apurada percepção auditiva para, na medida do possível, e seguindo o manual de transcrição, transpor para a escrita características da fala (pausas, hesitações, alongamentos de sons, ênfases, so-

breposição de falas etc), nem sempre de fácil notação, dadas as limitações do código escrito. Assim, na ausência do áudio correspondente, uma transcrição, quaisquer que sejam as notações sistematicamente empregadas, deve ser concebida tão somente como recurso auxiliar para o pesquisador recompor características da fala, por vezes indispensáveis ao entendimento de determinados fenômenos lingüísticos.

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, uma vez que:

> qualquer notação gráfica do oral é descontínua e dissociativa. Descontínua, pois tem de recorrer a *elementos discretos* (letras, palavras, frases), para representar o que se manifesta como um fluxo contínuo. Dissociativa, pois, por mais elaborado que seja, nenhum sistema de transcrição consegue reproduzir a *conjugação dos componentes segmental e suprasegmental* própria do discurso falado. (PAIVA, 2003, p. 135, grifos do autor)

Explicitada a natureza da transcrição, é necessário delimitar e justificar seu grau de detalhamento, cujas convenções influenciam a percepção dos dados lingüísticos (EDWARDS; LAMPERT, 1992). Portanto, a elaboração de um sistema de transcrição implica tomada de decisões teoricamente embasadas, que, uma vez registradas em manual próprio, orientam tanto os procedimentos para a transcrição das amostras de fala quanto a identificação de fenômenos lingüísticos de interesse do pesquisador.

Essas constituíram as principais motivações do Projeto ALIP, ao elaborar, conjuntamente com a equipe de transcritores, o seu manual de transcrição, ou seja, deixando claros os princípios organizadores do sistema de transcrição, já que subjaz a qualquer transcrição, mesmo que não explicitada, uma pré-análise de dados.

Dos procedimentos para a transcrição das amostras de fala

O sistema de transcrição do Projeto ALIP foi constituído a partir da comparação das convenções mais empregadas por outros projetos semelhantes: para o projeto NURC (CASTILHO, 1990); para o PEUL (PAIVA, 2003); para o VARSUL (VANDRESSEN, 1995) e para o projeto "Discurso & Gramática" (VOTRE; OLIVEIRA, 1995). Essas convenções foram, no entanto, aprimoradas, por meio da categorização das normas que as regulamentam, com o objetivo de explicitar a natureza do que se está sendo transcrito.

Dados os interesses do Projeto, para a transcrição ortográfica das amostras de fala foram adotas como categorias mais gerais as seguintes convenções:

(i) sobre grafia das palavras;

(ii) sobre alguns aspectos morfofonológicos;

(iii) sobre alguns elementos prosódicos;

(iii) sobre indicação de alguns aspectos da interação;

(iv) sobre comentários do transcritor.

Importante a destacar é que essas convenções foram estabelecidas à medida que o trabalho de transcrição ia sendo executado por uma equipe composta por 15 transcritores, número que se justificava pela quantidade de material sonoro a ser transcrito e o considerável dispêndio de tempo na execução dessa tarefa.[3] O sistema de transcrição em si, hoje em sua sexta versão, só se completou ao final dos trabalhos. Nesse ínterim, com o auxílio dos próprios transcritores, houve a oportunidade de se rever notações e de se acrescentar tantas outras. Esse procedimento, entretanto, exigiu, ao final dos trabalhos, uma revisão completa do material transcrito por uma equipe menor de trasncritores, quatro no total: três atuando na tarefa de revisão das transcrições e um, na tarefa de validação final das transcrições.

Acrescente-se que esse manual, mesmo em sua sexta versão, ainda está sujeito à revisão, considerando-se que qualquer que seja o sistema proposto será praticamente impossível dar conta do registro de todos os aspectos da fala. Entretanto, essa é a versão final que vem servindo de guia para a leitura e compreensão de todo material transcrito que compõe o banco de dados *Iboruna*.

Sobre as convenções adotadas

As convenções adotadas no sistema de transcrição do Projeto ALIP seguem detalhadas abaixo. Em cada quadro são apresentados as ocorrências, os sinais empregados em cada uma delas e uma exemplificação da notação.[4]

[3] A aplicação sistemática das convenções estabelecidas no manual de transcrição exigiu nada menos do que 12 horas de trabalho para a transcrição definitiva de uma hora de gravação, incluindo nesse tempo também as tarefas de correção e de revisão do material transcrito.

[4] Na coluna *Exemplos*, as expressões em negrito são apenas recursos para se colocar em destaque as notações exemplificadas. Tais recursos não fazem parte do sistema de transcrição.

OCORRÊNCIAS	SINAIS	EXEMPLOS
▪ Nomes próprios em geral.	Iniciais maiúsculas.	... O filme do **Almodóvar** ... (Não usar maiúsculas após os sinais "**...**" e "**?**")
▪ Nomes próprios que identificam o informante ou pessoa do relacionamento do informante.	Apenas as iniciais maiúsculas.[5]	Doc: Dona **M.**, a senhora falou que o **J.**, seu marido....
▪ Nomes de obras (livros, revistas, jornais) e palavras estrangeiras.	Em itálico, seguindo grafia da língua de origem.	... adorava ouvir ***Purple Rain***.. ... gostei de ler ***A viuvinha***...
▪ Marcadores discursivos.	Ocorrência seguida do ponto "?", quando for o caso.	... é pra deixar aqui **né**?... ... **então** acho que aí é o ponto...
▪ Interjeições dicionarizadas.	Ocorrência seguida do ponto "!".	**ah!** ... que alívio...**Vixe!**, **ixe!**, **pô!**,
▪ Numerais e letras.	Grafia por extenso.	... marquei com um **xix** a alternativa **bê** da questão **dois**...
▪ Siglas e abreviaturas.	Se pronunciada letra a letra, grafia em caixa alta, separando-se as letras por ponto.	**B.O.**, **I.N.S.S.**, **I.N.P.S.**, **U.F.R.J, R.G., C.P.F.**
	Se pronunciada como palavra, grafia prevista pela ortografia, em caixa alta e sem pontos entre as letras.	**USP**, **IAMSP**, **TAM**, **SUS**, **UFSCAR, CIC**.
▪ Redução de palavras.	Grafia da forma reduzida.	... Fiquei **deprê** com essa história.
▪ Truncamento (palavras incompletas).[6]	Emprego de barra após o truncamento.	... **ca/ casou** semana passada...
▪ Metalinguagem do informante.	Entre 'aspas simples'	... o '**mesmo'** do carioca...
▪ Citação.	Entre "aspas duplas"	... Armstrong disse **"pequeno passo para o homem... gigantesco salto para a humanidade"**

Quadro 1 – Sistema de transcrição do Projeto ALIP: convenções de grafia de palavras

As convenções adotadas para a grafia de palavras são, relativamente, as mais freqüentes nos sistemas de transcrição propostos para o Português Brasileiro, em razão 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,

[5] Esta convenção visa a manter o anonimato do informante. No arquivo sonoro, trechos correspondentes que identificam o informante foram "bipados", utilizando-se o programa *audacity 1.3*.

[6] Cabe observar aqui que o termo *truncamento* não se confunde com o termo *truncação*. *Truncação* (ou 'blend', ou 'palavra-portmanteau') denomina um tipo de composição vocabular, como por exemplo, *'portunhol'* de *'PORTUguês'* e *'espaNHOL'*. Truncamento indica a ocorrência de palavras incompletas ou também quando o falante detentor do turno é bruscamente interrompido pelo seu interlocutor.

estão arrolados truncamentos e metalinguagem do informante por contemplarem, em certa medida, aspectos da grafia do que é transcrito da língua oral.

OCORRÊNCIAS	SINAIS	EXEMPLOS
▪ Inserção de segmentos vocálicos.	Grafia da forma realizada.	**a**lembrá(r), **a**voá(r), d**r**ento, de**s**pois, revó**r**ve
▪ Apagamento de segmentos:	Segmento não realizado entre parênteses. Tonicidade da sílaba final de infinitivos marcada com acento agudo.	
a) segmento vocálico e/ou consonantal em início, meio ou final de palavra.		**(a)**rrancô**(u),** tam**(b)**ém, me**(s)**mo
b) ditongos.		c**a(i)**xa, p**e(i)**xe, p**o(u)**co
c) 's' do morfema de 1ª. pessoa plural.		nós fomo**(s)**, nós pegamo**(s)**
d) redução de gerúndio.		canta**n(d)o**, vive**n(d)**o
e) redução de infinitivo de verbos.		cant**á(r)**, vend**ê(r)**, sorr**í(r)**
▪ Uso de preposições:	Indicação da contração com apóstrofo.	
a) contração de *com* + artigo.		**c'a** (=com + a), **c'o** (=com + o), **c'um** (=com + um) **c'uma** (=com + uma)
b) contratação de *de* + artigo ou palavra iniciada por vogal.		**d'um** (=de + um), **d'uma** (=de + uma), **d'eu** (= de + eu), **d'oeste** (=de + Oeste), **d'água** (=de água), **d'onde** (de + onde).
c) contração/redução de *para* + artigo.	Grafia da forma realizada.	**pra** (sem acento), **pa** (sem acento), **pra** (=para + a), **pa** (=para + a), **pro** (=pra + o), **po** (=para + o), **pr´um(a)** (=pra + um(a)), **pum(a)** (=pa + um(a)).
d) modificação da preposição *em* (*em* > *ne*).		...a gente vai muito **ne** rio pa pescá(r)...
e) Inserção/modificação de preposição		eu penso **de** que ele é o melhor eu perguntei **na** onde ele ia eu perguntei **da** onde ele veio

Quadro 2 – Sistema de transcrição do Projeto ALIP: convenções de alguns aspectos morfofonológicos

Os aspectos morfofonológicos adotados no sistema de transcrição pretendem, de algum modo, preservar alguns aspectos do modo de realização da fala pelo informante. A representação de alguns outros aspectos pode esbarrar nas regras de ortografia que, "em razão de suas inúmeras arbitrariedades, podem falsear a realidade da variedade que se procura registrar" (PAIVA, 2003, p. 136), requerendo, portanto, outro sistema de transcrição que não o simplesmente ortográfico.

Em vista da grande massa de dados a ser transcrita e da impossibilidade do registro fiel e homogêneo de aspectos fonéticos mais detalhados da cadeia falada, a maioria dessas convenções se devem à opção de deixar transparecer, já nas transcrições, alguns fenômenos lingüísticos mais característicos da comunidade de fala.

OCORRÊNCIAS	SINAIS	EXEMPLOS
▪ Silabação.	Hífen entre as sílabas (sem espaço).	... foi quando ele disse...**fi-que-a-qui** ...
▪ Pausa (de qualquer extensão).	Reticências.	**...** ele**...** voltou feliz**...**
▪ Ênfase.	Em caixa alta.	... ele almoçou com **ELA**...
▪ Alongamento (vogais e consoantes).	Dois pontos digitados duas vezes.	... **ah::** ele **a::**cha...
▪ Interrogação.	Ponto de interrogação.	... você vai à festa**?**...

Quadro 3 – Sistema de transcrição do Projeto ALIP: convenções de alguns elementos prosódicos

A convenção para a notação de alguns elementos prosódicos trata particularmente de elementos suprassegmentais, como pausa, duração (alongamento de vogais e/ou consoantes), ritmo e velocidade de fala (silabação),[7] entoação (somente o padrão da interrogativa direta) e variação de altura e intensidade percebida como 'ênfase'.[8] O estabelecimento dessas convenções esbarra sempre numa análise mais apurada do dado a ser transcrito e, por isso, sujeitas a imprecisões por parte do transcritor, uma vez que se baseiam puramente na sua percepção auditiva.

A noção de *ênfase*, por exemplo, é bastante discutível e, conseqüentemente, a identificação de trechos de fala que sejam considerados enfáticos não é ponto pacífico. Diante da tarefa de transcrever *corpora* de fala, uma alternativa seria simplesmente não transcrever o que poderia ser considerado enfático. No entanto, a solução adotada para esse aspecto foi considerar

[7] 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*). (cf. MORAES; LEITE, 1993).

[8] Nos sistemas de transcrição consultados, a ênfase sempre é indicada, mas nunca explicitamente.

enfático somente os casos em que uma variação de altura – associada a uma ou mais sílabas (portanto pode ser parte de uma palavra, a palavra inteira) – for percebida com uma função de 'ênfase'. Desse modo, a ênfase realizada por meio da variação de altura será anotada a partir da percepção do transcritor. Ou seja, a notação terá como critério a percepção do falante de Português Brasileiro. Isso implica que interessados no tema 'ênfase' deverão empreender uma análise a partir de seus pressupostos teóricos.

Para além da ênfase, sobre cada um dos outros elementos prosódicos, várias observações poderiam ser feitas, mas apenas duas mais gerais parecem dar conta da opção feita. A primeira diz respeito à escolha desses e não de outros elementos prosódicos, opção que decorre da facilidade de sua percepção auditiva. A recorrência com que, por exemplo, a pausa é transcrita parece indiciar 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 notação da pausa num sistema de transcrição.

A segunda observação versa sobre o grau de detalhamento dos elementos prosódicos transcritos, os quais aparecem apenas anotados, no sistema de transcrição do Projeto ALIP, de modo a fornecer ao pesquisador interessado indícios para uma observação posterior mais refinada, com auxílio de ferramentas acústicas específicas, se for o caso. Ainda sobre a pausa, por exemplo, observa-se a opção por somente indicar a sua ocorrência, sem a preocupação com sua duração.

Justificada a escolha desses elementos prosódicos, uma reflexão sobre o desafio enfrentado quanto a esse aspecto da transcrição pode ser feita nos seguintes termos: seria suficientemente adequado propor que a seleção dos elementos prosódicos e seu grau de detalhamento estivessem baseados na relativa facilidade de percepção do transcritor, um falante nativo da variedade do português analisado?

A solução adotada inicialmente foi transcrever os elementos prosódicos mais freqüentemente anotados nos sistemas de transcrição disponíveis e com o mesmo grau de refinamento. Um cuidado extra consistiu na realização de um treinamento intensivo dos transcritores, a fim de explicitar técnicas de transcrição de base, de maneira que houvesse homogeneização do material transcrito.

OCORRÊNCIAS	SINAIS	EXEMPLOS
▪ Identificação dos participantes da interação.	Documentador (Doc.) Informante (Inf.) Interveniente (Int.)	**Doc.**: o senhor gosta de pescar? **Inf.**: eu não sei pescar... **Int.:** E aquele dia?
▪ Início de turno.	Em letras minúsculas.	Doc.: **o** senhor gosta de pescar?
▪ Discurso direto.	Travessão e aspas duplos.	... ela disse **– – "vamos à festa?" – –** eu respondi **– – talvez" – –**
▪ Seqüência de discurso direto.	Travessão e aspas duplos em cada um dos turnos, separados por reticências.	Inf.: aí ele falou **– – "cadê o dinheiro" – – ... – – "ta lá atrás" – –** o outro falou.
▪ 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. As sobreposições devem ser indicadas seqüencialmente ao longo da transcrição (1, 2, ..., n).	Inf.1: eu não tinha saído de lá... **[1][e foi então...]** Doc.: **[1][cê tava]** em casa **[2][ainda]**? Inf.1: **[2][eu tava].**.. aí ele ligou...
▪ Intervenção do documentador no fluxo de fala do informante.	Anotação no turno do informante, havendo ou não sobreposição de vozes.	Inf: outro dia eu estava na casa do João **[Doc.: ahan]** quando...
		Inf: outro dia eu estava na casa do **[1] [João] [1][Doc.: ahan]** quando ...
▪ Risadas simultâneas de documentador e informante.	Registro como comentário.	Doc e Inf: **((risos))**
▪ Marcadores de interação não-lexicalizados:	Grafia proposta, seguida de comentário, quando for o caso.	
a) concordância;		**Uhum / aham ((concordando))**
b) negação;		**hum hum / ham ham ((negando))**
c) manutenção do fluxo discursivo;		**hum/ham**
d) pergunta solicitando repetição.		**hum? / ham?**

Quadro 4 – Sistema de transcrição do Projeto ALIP: convenções de alguns aspectos da interação

Esse terceiro conjunto de convenções 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, indica-se também *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 e para qual são pertinentes duas questões: (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, esse aspecto, *mudança de seqüência temática*, foi descartado do sistema de transcrição do Projeto.

OCORRÊNCIAS	SINAIS	EXEMPLOS
▪ Comentário descritivo do transcritor.	Entre parênteses duplos.	... eu não gosto de pescar... é que não sei pescar... **((risos))** por isso que não gosto...
▪ Hipótese do que se ouviu.	Entre parênteses simples.	... foi então que ele (**fez**) a prova...
▪ Trecho ininteligível.	Registro por "Inint." entre parênteses	... foi então que ele (**inint.**) aí ele ...

Quadro 5 – Sistema de transcrição do Projeto ALIP: convenções de comentários do transcritor

A última categoria, *comentários do transcritor*, é um 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.

Sobre a validação das transcrições

Todas as gravações de AC e de AI, até o ponto definitivo de suas transcrições, foram submetidas a uma *primeira validação,* que exigiu de colaboradores mais experientes do Projeto a audição das amostras 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 textuais etc). A gravação invalidada foi alvo de nova coleta, com substituição de informante, quando necessário.

Uma *segunda validação* do material envolveu a checagem das transcrições no cotejo com as respectivas gravações e com o sistema de transcrição adotado. Essa tarefa foi executada por meio da troca das transcrições entre os transcritores. Aleatoriamente, os colaboradores do projeto também checavam os materiais validados nessa fase. A não correspondência entre gravação e transcrição implicou a refacção do trabalho.

Para garantir a homogeneidade das amostras definitivas, a fase final de validação do material transcrito foi executada por uma equipe de três membros, ficando a validação final de todas as transcrições a cargo de um quarto membro.

Das dificuldades encontradas na execução do Projeto

De importância central para esta exposição é o relato das dificuldades encontradas durante a constituição do banco de dados *Iboruna* e as soluções alcançadas com o desenvolvimento do projeto. Por uma questão de recorte, apenas duas dessas dificuldades serão detalhadas.

O uso dos gravadores digitais

Para evitar o uso de gravadores analógicos, foram empregados como recurso de gravação dos inquéritos gravadores digitais, com capacidade de armazenagem em memória de até 8 horas de gravações.[9]

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 pelo Projeto ALIP, em vista do alto custo desses gravadores à época, em média quatro vezes mais caros do que os gravadores que armazenam gravações em memória, e dos mini-discos.

Some-se a esse problema a constante necessidade de se apagar dos gravadores as entrevistas gravadas, após a sua transferência definitiva para as máquinas. Essa foi uma medida adotada, em vista do risco de se deixar uma gravação armazenada somente na memória do gravador até que se completassem todas as oito horas de gravação permitidas.

[9] Os gravadores utilizados são das marcas *gama-power* e *power-pack*, que vêm acompanhados de cabo de transmissão e de *software*, necessários para a conversão do arquivo armazenado na memória do gravador em arquivo de som no formato WAVE.

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.[10] 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, se comprada a gravações analógicas. Uma qualidade superior à obtida só seria possível se gravações fossem realizadas em laboratórios experimentais, o que não é o caso, tendo em vista a metodologia que o projeto emprega para a realização do censo lingüístico.

A etapa da transcrição das gravações

Para a etapa de transcrição das gravações, a meta esperada era a de que cada um dos 15 integrantes iniciais 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 demandava. Além disso, o treinamento da equipe 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 ainda não previstas no manual de transcrição. Todas as transcrições foram realizadas usando-se os recursos disponíveis do *media player*.

Como somente com o avançar do trabalho de transcrição foi possível medir a real dimensão dessa tarefa, na etapa final do projeto foi designada uma equipe menor de revisores das transcrições, com tarefa exclusiva de se checar a aplicação das convenções do manual de transcrição, de modo a se garantir uma homogeneidade mínima do material transcrito.

Parece, hoje, ingenuidade pensar em homogeneidade de procedimentos, quando se lida com uma equipe grande de pessoas. 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.

[10] 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.

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. Citem-se, como principais problemas, a invalidação de gravações resultantes da desatenção do documentador na condução das entrevistas e a revisão de transcrições totalmente inadequadas ao sistema proposto.

Uma primeira medida para contornar esses problemas foi a finalização do manual de transcrição e a elaboração de manuais de todas as etapas do projeto. Outra solução foi a substituição de alguns membros da equipe técnica e, na fase final do projeto, a redução da equipe responsável pelo material transcrito.

Importância da transcrição para análise de fenômenos lingüísticos

Em vista das convenções adotadas no manual de transcrição, relativamente à dispensabilidade ou não do recurso ao áudio, as pesquisas que vêm se valendo das transcrições podem ser enquadradas em dois grupos.

Dada a natureza dos fenômenos investigados, o primeiro grupo dispensaria, *a priori*, da parte do analista, a audição da gravação, como é o caso da descrição dos seguintes fenômenos:

- concordância nominal (FIAMENGHI, 2007; SALOMÃO, 2008);
- concordância verbal (RUBIO, 2008);
- uso alternante de indicativo/subjuntivo (SANTOS, 2005);
- redução de gerúndio (FERREIRA, 2008);
- modalidade/evidencialidade (VENDRAME, 2006; HENGEVELD et al., 2007);
- articulação de orações (SANTANA, 2005; GONÇALVES, 2008; FORTILI, 2007).

Fenômenos como esses, cuja notação está devidamente prevista, requerem sempre o cotejo da transcrição com a gravação, para confirmação ou não de sua realização, não tanto pela falta de aplicação das notações convencionalizadas, mas mais pelo cuidado por parte do analista, uma vez que o registro, por vezes, pode ter escapado à percepção auditiva do transcritor, principalmente nos casos de fenômenos que envolvem variações morfofonológicas.

Integrantes do segundo grupo, outros fenômenos, porém, não partilham dessa mesma "facilidade, dadas as restrições do código ortográfico frente ao registro de questões de ordem fonológica e, principalmente", de ordem prosódica, que são, por

vezes, decisivas para o analista no enquadramento do dado que analisa nessa ou naquela categoria. São exemplos de fenômenos lingüísticos, cujas peculiaridades não previstas no sistema de transcrição do projeto ALIP, tornam indispensáveis a audição das gravações e, por vezes, até o recurso a ferramentas de análise acústica mais apurada:

- o alçamento de vogais pretônicas e postônicas, na caracterização do sistema vocálico do dialeto riopretano, como em *menino* / *mininu* (SILVEIRA, 2008; CARMO, 2008; RAMOS, 2008);
- a haplologia, como em *faculdade de letras*/*faculdade letras* (PAVEZI, 2006);
- o uso de marcadores discursivos e sua incidência sobre segmentos tópicos (GUERRA, 2007; LOPES, 2008; PENHAVEL, 2005).
- a gramaticalização de itens/construções lexicais, com conseqüente perda/fusão de material fonológico, em que se modificam as fronteiras de constituintes (GALBIATI, 2008; FELÍCIO, 2008; ROSA, 2005; CINTRA, 2006).

Considerações finais

Embora não trabalhando muito próximo das 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 da pesquisa lingüística realizada no interior do Estado de São Paulo.

Diante do propósito desse 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. Encontra-se também à disposição, gratuitamente, no *site* do projeto (http://www.iboruna.ibilce.unesp.br), parte do material integrante do banco de dados *Iboruna*, inclusive o material sonoro. Em vista dos recursos públicos investidos na sua constituição, sem o qual, certamente, o projeto não seria executado, em breve todo o banco de dados estará disponível.

Esperamos, assim, que, ao mesmo que sirva de instrumento de divulgação, esse artigo sirva também de guia prático na execução de projetos semelhantes, principalmente no que tange a mais árdua tarefa de constituição de banco de dados com amostra de fala: a etapa de transcrição.

Abstract

In this paper, we present methodological procedures to transcription of speech samples integrating the ALIP (Linguistic Sample of the Inner Cities of São Paulo State Project). We justify the conventions adopted in a transcription system, present some problems of the transcription work and their solutions, and exemplify linguistic phenomena that we have studied based on the transcriptions of speech samples.

Keywords: *Transcriptio. Speech sample. Database.*

Referências

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

CINTRA, M. R. *Gramaticalização da perífrase ir + estar + gerúndio*. Programa de Pós-graduação em Lingüística do IEL/UNICAMP. Bolsa FAPESP. Orientador: I.G.V. Koch. Em andamento, a partir de março/2006.

EDWARDS, J. A.; LAMPERT, M. D. (Ed.). *Talking data*: transcription and coding in discourse research. Hillsdale: Lawrence Erlbaum Associates, 1992. p. 3-31.

FELICIO, C. P. *Gramaticalização da conjunção concessiva "embora"*. 135f. Dissertação (Mestrado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2008. Orientador: S.R.Longhin-Thomazi.

FERREIRA, J. S. *O gerúndio na fala do interior paulista*. Iniciação Científica. Departamento de Estudos Lingüísticos e Literários, Universidade Estadual Paulista, São José do Rio Preto. 2007. Orientador: L.E. Tenani.

FIAMENGHI, A. H. R. *Motivações formais da marcação de pluralidade no SN na função de sujeito e complemento*. Iniciação Científica. FAPESP, 2007. Orientador: R.G. Camacho.

FORTILI, S. *As construções não-verbais no português falado no interior do Estado de São Paulo*. 120f. Dissertação (Mestrado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2007. Orientador: E.G.Pezatti.

GALBIATI, M. E. *Gramaticalização das perífrases verbais "agora que" e "já que"*. 160f. Dissertação (Mestrado em Lingüística e Língua Portuguesa)-Universidade Estadual Paulista, Araraquara, 2008. Orientador: Maria Helena de Moura Neves.

GONÇALVES, S. C. L. *Aspectos da subordinação sentencial sob uma perspectiva diacrônica: o caso das orações em posição argumental de sujeito*. Trabalho Submetido à publicação no livro do "Projeto Para História Português Paulista". 2008.

GUERRA, A. R. *Funções textual-interativas dos marcadores discursivos*. 181f. Dissertação (Mestrado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2007. Orientador: S.C.L. Gonçalves,

HENGEVELD, K. The architecture of a functional discourse grammar. In: MACKENZIE, J. L.; GÓMEZ GONZÁLES, M. A. (Ed.). *A new architecture for functional grammar*. Berlin: Mouton de Gruyter, 2002.

HENGEVELD, K. et al. The expressibility of modality in representational complement clauses. *Alfa*: Revista de Lingüística, São Paulo, v. 52, 2007. Disponível em:<http://www.alfa.ibilce.unesp.br>.

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

LOPES, L. R. *A emergência do marcador discursivo "assim" sob a óptica da gramaticalização*. 240f. Dissertação (Mestrado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2008. Orientador: Sanderléia Roberta Longhin-Thomazi.

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, SP: Ed. da Unicamp, 1993. v. 2, p. 67-77.

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

PAVEZI, V. C. *A haplologia no dialeto paulista*. 126f. Dissertação (Mestrado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2006. Orientador: L.E. Tenani.

RAMOS, A. P. *As vogais postônicas mediais no dialeto paulista*. Dissertação (Mestrado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2008. Orientador: L.E.Tenani. Em andamento.

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

ROSA, E. F. *Os advérbios de tempo e lugar no português brasileiro*: casos de gramaticalização ou de discursivização? Tese (Doutorado em Lingüística)-Universidade Estadual de Campinas, Campinas, SP, 2005. Orientador: I.G.V.Koch. Em andamento

RUBIO, C. F. A concordância verbal de 3ª. pessoa na fala da região noroeste do Estado de São Paulo. 158f. Dissertação (Mestrado em

Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2008. Orientador: S.C.L. Gonçalves.

SALOMÃO, M. H. *A variação de pluralidade no SN-predicativo na variedade falada na região de São José do Rio Preto.* Dissertação (Mestrado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2008. Orientador: R.G. Camacho. Em andamento.

SANTANA, L. *Motivações funcionais da gradação entre construções encaixadas nominais e verbais*. Doutorado (Doutorado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2005. Orientador: R.G. Camacho. Em andamento.

SANTOS, R. M. A. *O uso variável do modo subjuntivo em estruturas complexas*. 150f. Dissertação (Mestrado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2005. Orientador: S.C.L. Gonçalves.

SILVEIRA, A. A. M. *As vogais pretônicas na fala culta do interior paulista*. 148f. Dissertação (Mestrado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio Preto, 2008. Orientador: L.E.Tenani.

SOUZA, E. P. *A função dos marcadores discursivos na estruturação do discurso*. Tese (Doutorado em Lingüística)-Universidade Estadual de Campinas, Campinas, SP, 2005. Orientador: I.G.V. Koch. Em andamento.

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. In: ENCONTRO NACIONAL SOBRE LÍNGUA FALADA E ENSINO, 1., 1995, Maceió. *Anais*... Maceió: EDUFAL, 1995. p. 196-221.

VENDRAME, V. *O lugar da inferência no sistema evidencial da língua portuguesa*. Tese (Doutorado em Estudos Lingüísticos)-Universidade Estadual Paulista, São José do Rio, 2006. Preto. Orientador: M.M.Dall´Aglio-Hattnher. Em andamento.

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.