Banco de dados orais: uma nova perspectiva aos estudos do português brasileiro

Regina Célia Fernandes Cruz
Universidade Federal do Pará

Jailma do Socorro Uchôa Bulhões
Bolsista – IC – UFPA

Léa da Silva Fernandes
Bolsista – IC – UFPA

Abstract

This work represents an attempt to facilitate the automatic access to the spontaneous Portuguese data for studies on experimental phonetic area and speech synthesis, maintaining qualitative and quantitative information of the data. This corpus constitutes more than 30 hours of recording with samples of two Portuguese spoken dialects in Amazonia Region: the Amazon Portuguese (AM) and the Afro-Brazilian Portuguese (ABP). As a baseline, samples of the Standard Brazilian Portuguese (PB) were also used. The corpus inventory, the digitalization of the speech data and the phonetic segmentation using the PRAAT software preceded the construction of the oral database. Using language HTML and other resources of computer science a documentation of the database in a Internet site becomes possible.

1. INTRODUÇÃO

O objetivo principal do presente estudo é o de demonstrar a importância de utilização de dados espontâneos nos estudos acústico-experimentais, assim como o da conciliação de *corpus* sociolingüísticos e da metodologia fonético-experimental.

A proposta de trabalho aqui descrita nasceu das dificuldades encontradas por Cruz (2000b) na exploração da totalidade de seu *corpus* e que impossibilitaram um aprofundamento nos processos estudados. Trata-se, portanto, da atualização de um banco de dados orais concebido e já executado desde Cruz (2000b). A organização e disponibilidade desse *corpus* num banco de dados destaca-se na sua originalidade por ser uma base de dados orais constituída de fala espontânea, útil para estudos de variação lingüística do português.

A necessidade de documentar e disponibilizar eletronicamente dados de fala espontânea representativos do português brasileiro decorre da quase total inexistência de *corpora* organizados em forma de banco de dados orais e disponíveis eletronicamente a fim de servir de suporte para toda e qualquer investigação lingüística, assim como para aplicação tecnológica envolvendo engenharia de fala.

O crescente interesse dos estudos fonéticos em relação à fala espontânea não é suficiente para uma aplicação imediata e direta da metodologia da fonética experimental a dados de fala espontânea. A impossibilidade maior está na forte tradição de utilização da fala lida ou paradigmaticamente gerada nas experimentações em fonética refletida, por exemplo, na alta sensibilidade dos equipamentos utilizados nesse tipo de estudo.

No sentido de esclarecer melhor tais dificuldades, uma descrição detalhada das mesmas é feita ao longo do presente estudo, assim como dos procedimentos empreendidos na tentativa de facilitar o

acesso automático a dados do português espontâneo por estudos na área de fonética experimental e engenharia de fala, através da organização de um *corpus* em banco de dados orais, contendo informações qualitativas e quantitativas dos sinais sonoros armazenados. Qual então é o *corpus* em questão?

2. CORPUS

O referido *corpus* compreende mais de 30 horas de gravação com amostras de três variedades lingüísticas do português: (i) o português regional paraense (AM), coletado no período de 1989 –1990 (TRINDADE); (ii) o português afro-brasileiro (ABP), variedade alvo, falada pelas comunidades quilombolas do Pará, coletado entre 1992-1996 (CRUZ, 2000b) e; (iii) o português brasileiro padrão (PB) emprestado de Oliveira (2000), com amostra da variedade lingüística em grandes centros urbanos e por falantes de alto nível de escolaridade.

Dois aspectos os aproximam:

a) a língua: todos contêm amostras do português brasileiro;
b) a metodologia seguida para sua formação: os três *corpora* foram formados prioritariamente para servir de suporte empírico a investigações sociolingüísticas.

Inicialmente, todos os sinais sonoros armazenados nesse banco de dados orais foram coletados em fitas cassete em trabalho de campo. Das quais foram retiradas as informações necessárias acerca do *corpus*. Antes da digitalização dos sinais sonoros, Cruz (2000b) procedeu a um inventário das cassetes de áudio. Esse inventário permitiu identificar quais seriam as gravações mais apropriadas para os estudos fonéticos. A escuta permitiu chegar a uma classificação das fitas em três categorias indicadas através de cores. Essa classificação repousa sobre a qualidade sonora das gravações que varia de uma tomada de gravação[1] a outra.

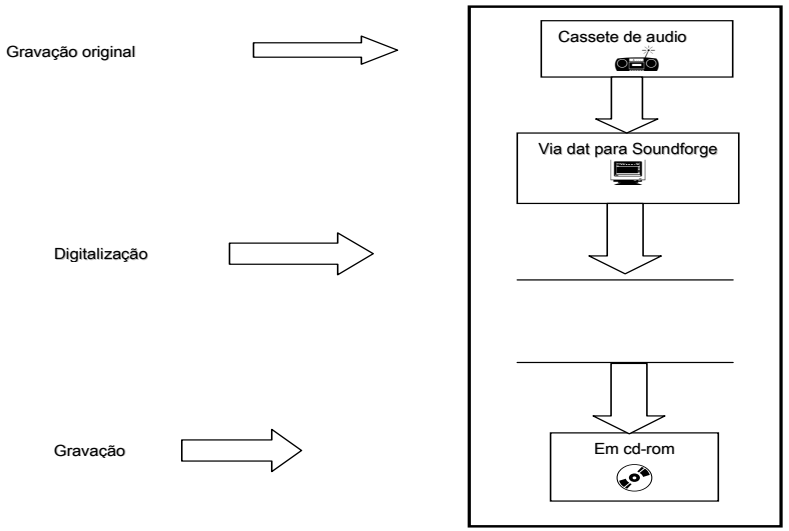
Esta mesma classificação é retomada em um arquivo específico do banco de dados orais aqui descrito. As tomadas de gravação

foram então identificadas como: <<azul>> (tomadas de gravação de excelente qualidade sonora); <<amarelo>> (tomadas de gravação cuja utilização para análises acústicas é duvidosa) e <<vermelho>> (tomadas de gravação cuja situação de fala não faz parte daquelas consideradas apropriadas para um estudo sobre fala espontânea ou tomadas de gravação inutilizáveis para um estudo acústico). Além de precisar sobre a qualidade das gravações, o inventário feito por Cruz (2000b) também procedeu a um levantamento acurado sobre faixa etária, sexo e procedência dos locutores, situação de fala, local das gravações e variedade lingüística. Todas essas informações figuram no código atribuído a cada sinal sonoro no momento da digitalização.

3. DIGITALIZAÇÃO

Uma exigência para facilitar o tratamento experimental foi a transferência do *corpus* para uma base fixa. Para esta fase, também o inventário preliminar do *corpus* foi fundamental. Para a transferência do *corpus* para CDROM, Cruz (2000b) decidiu fazer a transferência de todas as fitas cassetes classificadas como AZUL e AMARELO, bem como as gravações classificadas como VERMELHO, que não apresentavam problemas de rotação da gravação.

A digitalização das gravações em sinais de áudio foi feita no programa SOUNDFORGE via um gravador DAT. Os parâmetros utilizados por Cruz (2000b) foram: 44100 Hz, 16 bit, Mono. A saturação do sinal foi controlada por um monitor de maneira que o TBS áudio remixador permitisse verificar se a digitalização estava correta. O esquema 1 explica a maneira como foi feita a digitalização.



Esquema 1 – Fases do processo de digitalização das gravações
(extraído e traduzido de CRUZ 2000b: 183)

O resultado desta primeira fase foi a gravação de 38 CD-ROMS áudio que totalizaram 39 horas, 19 minutos e 55 segundos.

A fase de digitalização foi extremamente importante para a organização dos dados. A unidade "tomada de gravação" estabelecida como informação pertinente para o *corpus* foi retomada nesta fase. A noção de tomada de gravação aparece pela primeira vez no momento de preparar o inventário do *corpus*: o critério utilizado para a delimitação destas tomadas de gravação é simplesmente o fato de ligar e desligar o gravador. Por esta razão, «tomada de gravação» foi empregada com o sentido de «evento sonoro gravado sem interrupção». Elas são a unidade base que compõe os sinais sonoros armazenados

O programa SOUNDFORGE permitiu reunir duas ou mais tomadas de gravação que pertenciam a uma mesma situação de fala e que, por razões técnicas (limitação física do tempo da fita cassete por exemplo), se encontravam em tomadas de gravação separadas. No caso de uma mesma fita cassete conter várias situações de fala diferentes, as quais nem apresentavam relação direta, foi possível a

separação das mesmas; assim como foi possível também reunir em um mesmo sinal, por exemplo, todas as tomadas de gravação que pertenciam à fala lida. Estas foram reunidas em um mesmo sinal e separadas por um intervalo de silêncio de 600 milissegundos. Foi igualmente possível eliminar barulhos indesejáveis para análise acústica.

Para não se perder a vinculação com as gravações originais, cada sinal gravado contém as seguintes informações: um número de ordem, a data da gravação e a duração total do sinal de áudio que ele contém, além de detalhes de cada arquivo digitalizado por SOUNDFORGE (assinalando data da digitalização, as tomadas de gravação originais, e a duração do sinal .WAV obtido). As outras informações estão disponíveis no próprio código atribuído a cada sinal .WAV, o qual indica: a situação de fala, sexo do locutor, lugar e qualidade da gravação.

A digitalização com SOUNDFORGE e a transferência para CDROM permitiu organizar melhor os sinais sonoros, entretanto, estes ainda não estavam prontos para o tratamento com os programas de análise acústica experimental. Era necessário que os CDS áudio fossem transformados em CDS de dados e para essa tarefa foi utilizado o programa COOLEDIT. Cruz (2000b) fez a redigitalização dos sinais sonoros de acordo com os seguintes parâmetros: freqüência de 16kHz, 16 bits e sinal mono. Desta forma, o COOLEDIT fornecia um sinal .WAV pronto para ser trabalhado com os programas de tratamento acústico. Nem todos os sinais áudio de 44100 hz foram redigitalizados em 16 kHz. A autora selecionou apenas aqueles contendo mais amostras de fala espontânea, assim como os sinais que apresentavam melhor qualidade de gravação. Por esta razão, ela escolheu, em um primeiro momento, 61 sinais de áudio a serem redigitalizados.

Houve também a redigitalização de 7 sinais não pertencentes a fala espontânea (1 de celebração religiosa e 6 de fala lida) com o objetivo de estabelecer comparações futuras com outras situações de fala. No total, 5 CDs de dados foram gravados.

Todo este trabalho de transferência dos dados para CDROM permitiu trabalhar os sinais tais quais se apresentavam nos dados em

tomadas de gravação, evitando, assim, a digitalização direta nos programas de análise acústica, pois mesmo o PRAAT, que permite analisar sinais de mais de 3 horas de duração, restringe a duração a um minuto por aquisição de sinal. Mesma situação Cruz (2000b) enfrentou com o programa MES, que permitia apenas um minuto de aquisição direta. Desta maneira, no caso do PRAAT foi possível fazer a leitura do sinal diretamente do CDROM.

Se um *corpus*, que não foi concebido inicialmente para sustentar um estudo experimental, mas com a intenção de ser representativo da variedade lingüística estudada, permite fazer análises acústicas, é possível que esta proporção aumente ainda mais em um *corpus* formado com a intenção, desde o início, de se obter o espontâneo para estudos experimentais. O *corpus* do PB que foi emprestado é um exemplo disso, pois o autor formou um *corpus* representativo do PB com uma alta fidelidade de gravação (OLIVEIRA, 2000).

A etapa de digitalização foi um trabalho essencial para a construção desse banco de dados, já que permitiu obter sinais sonoros prontos para serem tratados em programas de análise acústica. O intuito não é apenas manter um acervo documentado do português brasileiro, interessa também tornar o acesso a esses dados facilitado. E a forma mais viável encontrada foi a organização do *corpus* em um banco de dados orais, o qual será descrito a seguir.

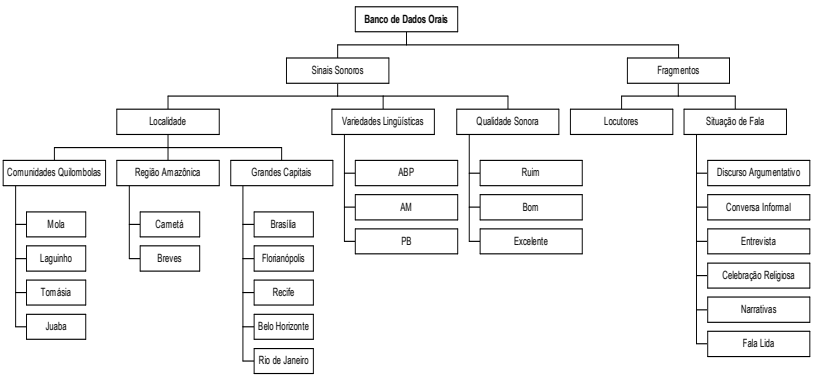
4. DESCRIÇÃO DO BANCO DE DADOS

Como se tratava de um *corpus* coletado em trabalho de campo, antes da concepção e posterior execução do banco de dados, como já mencionado acima, foi feito um inventário do material lingüístico presente no mesmo (CRUZ, 2000b). A informação levantada: número de locutores, a diversidade de situação de fala, tempo, localidades e datas das gravações foram importantes para a construção do banco de dados, que contém informações tanto qualitativas, quanto quantitativas sobre as gravações. Vejamos como estas informações estão disponíveis no banco de dados concebido para permitir o manuseio automático desse material lingüístico.

O banco de dados é composto de sete arquivos interligados: (i) o das localidades, contendo informações sobre os povoados e as cidades de procedência dos locutores; (ii) o dos locutores; (iii) o dos sinais sonoros, contendo informações diretas sobre as tomadas de gravações e; (iv) um arquivo chamado fragmento. Este último arquivo contém as informações qualitativas (características do locutor, situação de fala, e também transcrição fonética e lexical do sinal sonoro) e informações quantitativas (medidas físicas sobre gravações, as segmentações realizadas, o sinal sonoro); (v) o das variedades lingüísticas; (vi) o da qualidade dos sinais sonoros; (vii) o da situação de fala.

Por se tratar de um banco de dados do tipo relacional, inicialmente foi proposto a utilização do programa FILEMAKER PRO.4 na execução do banco de dados orais (CRUZ, 2000b; CRUZ, HIRST & BEL). Todavia, constatando uma incompatibilidade entre o FILEMAKER PRO.4 e o programa PRAAT, utilizado para a segmentação dos sinais sonoros em palavras e fonemas, preferiu-se tentar a edição de um *site* na *Internet,* usando linguagem HTML e outros recursos de animação sonora.

O esquema 2 mostra como o banco de dados orais está estruturado hierarquicamente.



Esquema 2 – Estrutura hierárquica do banco de dados

Como podemos perceber, este banco de dados é composto por arquivos interligados que apresentam informações tanto qualitativas, quanto quantitativas dos sinais sonoros. Além da página inicial, o banco de dados orais é composto dos seguintes arquivos:

4.1. Sinais Sonoros

Este arquivo contém informações diretas sobre os sinais sonoros, que correspondem à identificação do locutor, à duração do sinal sonoro, à situação de fala, à data, ao lugar e à qualidade de gravação. Assim como disponibiliza uma segmentação em turnos de fala de cada sinal sonoro e possibilita uma comprovação auditiva desta mesma segmentação em um arquivo .mp3. Estes sinais sonoros correspondem à unidade tomada de gravação descrita no item 2.

O gráfico 1 mostra a porcentagem de sinais sonoros armazenados já transcritos em turnos de fala.

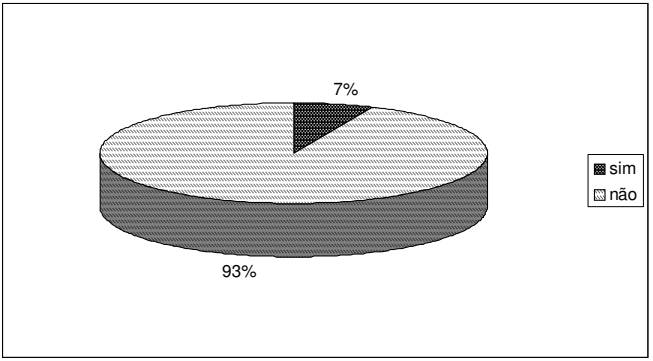


Gráfico 1 – porcentagem de sinais sonoros armazenados já transcritos em turno de fala (total de 61 sinais sonoros).

Para figurar no banco de dados, os sinais sonoros foram codificados segundo um padrão de notação cuja convenção está descrita no quadro 1 e na figura 1 abaixo:

Variedade lingüística	Local	N.º	Tomada	Gravação	Situação de fala	Qualidade da gravação
A: Afro	**M**: Mola				**E:** entrevista	**E**: excelente
Am: Amazonia	**J:** Juaba				**CI**: conversa	(= azul)
B: Brasileiro	**T:** Tomásia				informal	**B**: bom
P: Português	**L:** Laguinho				**N:** narrativa	(= amarelo)
	C: Cametá				**CR**: Celebração	**R**: ruim
	Cv: Custavê				religiosa	(= vermelho)
	A: Ajó				**TA:** discurso	
	I: Itapocu				argumentativo	
	BR: Brasília				**L**: fala lida	
	CR: Curitiba				**X:** não identificado	
	RJ: Rio de Janeiro					
	BH: Belo Horizonte					
	FL: Florianópolis					

Quadro 1 – todas as informações utilizadas na notação dos sinais sonoros (emprestado e traduzido de CRUZ, 2000b: 267)

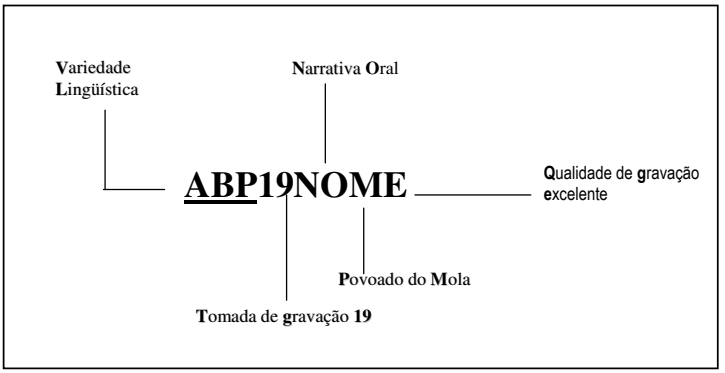


Figura 1 – tomadas de gravação foram codificadas segundo as seguintes convenções (emprestado e traduzido de CRUZ, 2000b: 291): *ABP19NOME* = Variedade lingüística (*ABP*); tomada de gravação (*19*); situação de fala (*narrativa oral*); localidade (*povoado do Mola*); qualidade de gravação (*Excelente - E*).

O arquivo de sinais sonoros contém as seguintes informações:

4.1.1 Localidades

Este arquivo contém prioritariamente informações sócio-econômicas e histórico-culturais das comunidades quilombolas de Juaba, Mola, Tomásia e Laguinho, localizadas no município de Cametá (PA). No banco de dados há referência tanto à localidade de origem dos locutores, quanto às localidades onde foram feitas as gravações.

4.1.2 Variedades Lingüísticas

Este arquivo contém informações sobre as três variedades do português falado no Brasil com amostras no banco de dados.

a) AM – português regional da Amazônia [TRINDADE]. Foi também o *corpus* utilizado para as primeiras experimentações com fala espontânea;

b) PB – português brasileiro padrão. *Corpus* complementar emprestado de Oliveira [7] para efeito de comparação;

c) ABP – português afro-brasileiro como definido por Cruz (2000b).

O gráfico 2 apresenta como as variedades estão distribuídas no banco de dados em número e duração de sinais sonoros.

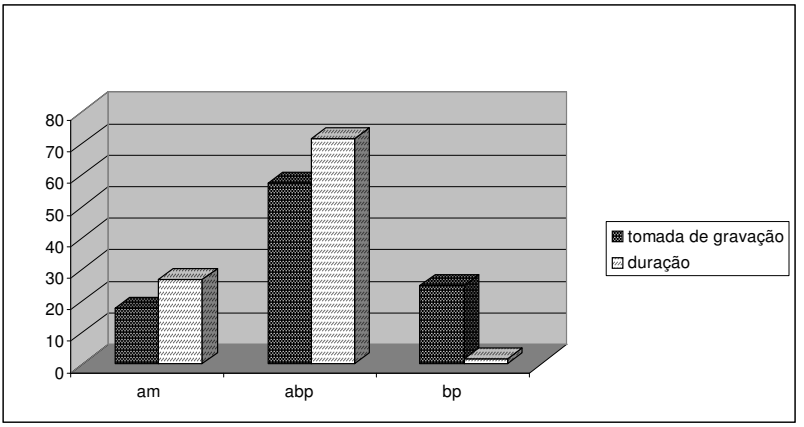


Gráfico 2 – visão do *corpus* em variedades lingüísticas, tomada e duração de gravação (total de 61 sinais sonoros e duração total de 69.760 segundos) (emprestado e traduzido de CRUZ 2000b:185)

Os *corpora* foram formados em momentos distintos. O português regional da Amazônia foi o primeiro a ser formado, sua coleta de dados se deu entre setembro de 1989 a fevereiro de 1990 [8]. O português afro-brasileiro foi coletado durante três trabalhos de campo: (i) novembro a dezembro de 1992; (ii) outubro de 1993 a janeiro de 1994 e (iv) setembro a dezembro de 1994 (CRUZ, 2000b). Os dados do português brasileiro padrão foram gravados em 1998 com falantes do português residentes em Vancouver no Canadá (OLIVEIRA, 2000), todos de classe média alta e com nível de escolaridade alto. Foram dados coletados com o objetivo de se estudar a fala espontânea do ponto de vista experimental e numa visão laboviana de trabalho de campo.

Este arquivo apresenta informações sociolingüísticas e lingüísticas sobre as referidas variedades, assim como os resultados de Cruz (2000b) e Oliveira (2000).

4.1.3. Qualidade Sonora

Descreve-se como as tomadas de gravações foram transformadas em sinais sonoros, tomando o critério da qualidade das gravações, e indica quais são os sinais sonoros de *excelente qualidade de gravação* e quais *não são aconselháveis* para um estudo acústico-experimental.

O gráfico 3 mostra o *corpus* original ainda em cassete, e o gráfico 4 mostra o *corpus* final digitalizado. Ressaltando que, no gráfico 3, os 20% de gravação considerados ruim indicam tanto problemas de ordem técnica (rotação irregular, muito barulho), quanto dados não representativos de fala espontânea. Enquanto o gráfico 4 indica que 7% das gravações etiquetadas como VERMELHO são amostras de fala não-espontânea.

Apesar de se tratar de um *corpus* não concebido especificamente para sustentar um estudo fonético experimental, obtiveram-se tomadas de gravação passíveis de serem utilizadas nesse tipo de estudo. Logo, essa proporção deve aumentar em coleta de dados, visando obter tanto fala espontânea, quanto gravações com alta fidelidade.

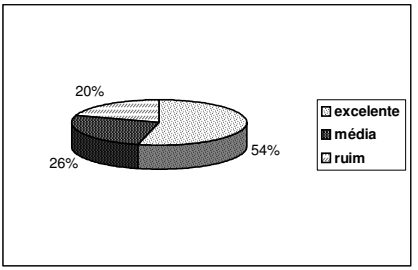


Gráfico 3 - *corpus* original ainda em cassete

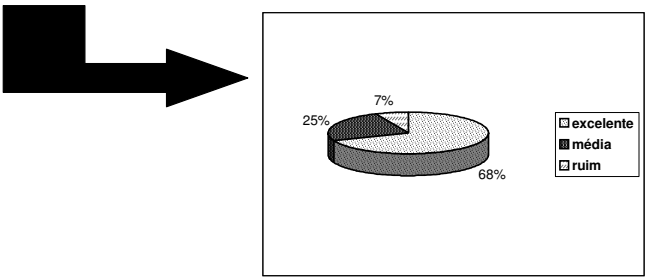


Gráfico 4 - *corpus* final digitalizado

4.2. Fragmento

Este arquivo é o mais importante para quem está realizando estudos acústicos, pois contém informações de acesso aos segmentos sonoros do *corpus* armazenado. Aqui um fragmento sonoro corresponde a um turno de fala, um só enunciado de um único locutor. Os resultados das análises acústicas, como: duração, intensidade e freqüência, estão estocadas neste arquivo. Em termos de tamanho, é o arquivo mais volumoso, pois ele contém um grande número de fichas. As fichas contêm as informações que permitem o acesso ao sinal acústico de formato .mp3, e a sua segmentação em palavras e fonemas. No que diz respeito à segmentação feita pelo alinhador do MBROLIGN, estas foram convertidas para o formato TEXGRID, aceito por PRAAT, através de *script* PERL e sofreram revisão manual.

Além do sinal sonoro em si, nós podemos acessar as informações contextuais mais precisas sobre os sinais sonoros tratados de forma automática.

Os fragmentos sonoros são transcritos lexical e foneticamente, com os quais é possível realizar as análises prosódicas de ritmo e entoação, pois sua unidade de base compreende a voz de único locutor. Eles estão sendo transcritos com PRAAT para a parte segmental e a codificação MOMEL para a normalização da curva de F0 (CRUZ, 2000b). A transcrição segmental é feita com o alfabeto SAMPA. Há 80 fragmentos sonoros segmentados até o momento.

Este arquivo contém os seguinte sub-arquivos:

4.2.1. Locutores

Este arquivo apresenta informações sobre sexo, faixa etária, origem, escolaridade e variedade lingüística dos locutores.

Os gráficos abaixo mostram informações dos locutores quanto ao sexo e à faixa etária. O gráfico 6 especifica que, no *corpus* há um número maior de vozes masculinas. E o gráfico 7 mostra que há uma maior representatividade de locutores na faixa de 30-70 anos de idade.

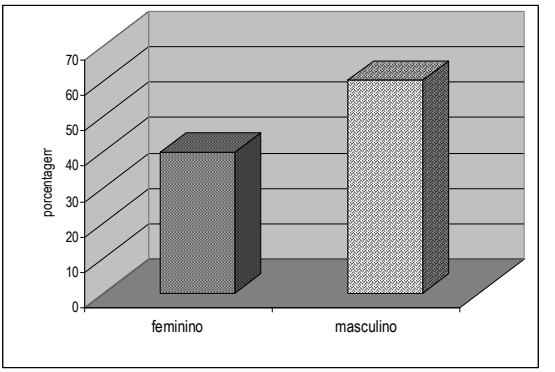


Gráfico 5 - informações dos locutores quanto ao sexo

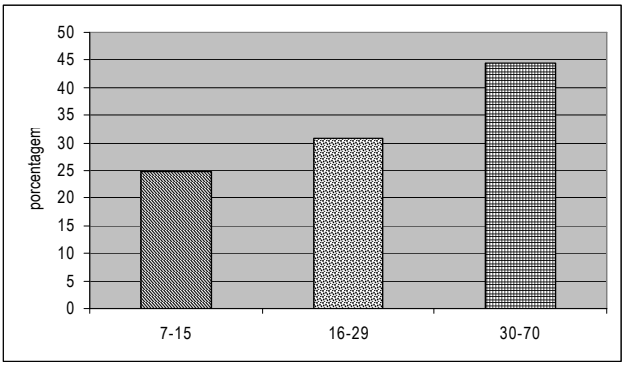


Gráfico 6 - distribuição dos locutores por faixa etária

4.2.2 Situação de Fala

Apresenta a distribuição dos 61 sinais sonoros em situação de fala, como: discurso argumentativo, conversa informal, entrevista, celebração religiosa, narrativas e fala lida.

Foram escolhidos, em um primeiro momento, 61 sinais de áudio, distribuídos da seguinte forma como expresso no gráfico 7.

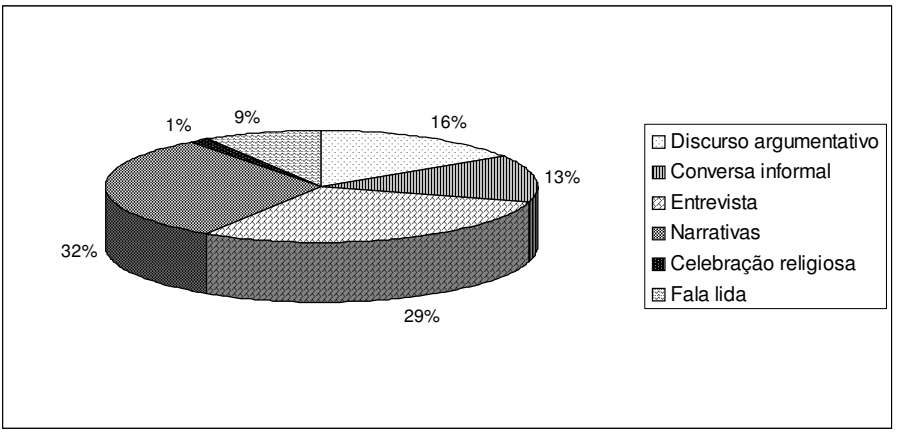


Gráfico 7 - composição dos Cds de dados em situação de fala

5. COMO INTERROGAR O BANCO DE DADOS?

Não há uma linguagem de interrogação específica para utilizar este banco de dados, pois os operadores lógicos necessários a uma busca estão disponíveis diretamente na interface gráfica do *site Internet*.

Este banco de dados apresenta informações diretas sobre os sinais sonoros, que correspondem a: identificação do locutor, data e lugar de gravação, qualidade da gravação, situação de fala, e duração do sinal. Encontra-se disponível, ainda, uma segmentação em turnos de fala de cada sinal sonoro, contendo as medidas de duração e freqüência referentes a cada segmento em um arquivo ASCII. Também é possível uma comprovação auditiva desta mesma segmentação em um arquivo mp3.

6. DESCRIÇÃO DO SITE

O usuário do banco de dados poderá acessá-lo através do *site* que o hospeda e que possui, como "porta de entrada" e página principal, uma *Home Page,* contendo a página principal do *site,* e como tal funciona como portal para todas as outras páginas, permitindo assim que estas estejam interligadas entre si. Em sua totalidade o *site* apresenta 23 documentos em HTML[2] que apresentam descrições detalhadas sobre o trabalho desenvolvido no projeto, sobre os integrantes da equipe de trabalho, assim como dos colaboradores e consultores científicos. Também apresenta uma página com formulário que permite ao cliente (usuário) interagir com o servidor, preenchendo campos, clicando em botões e passando informações, o que possibilita ao visitante entrar em contato com a coordenação do projeto e emitir suas opiniões. E tem como documento principal uma página intitulada Banco de Dados Orais, que funciona como um *link* que interliga a *Home Page* ao banco de dados orais, permitindo, assim, uma perfeita interação entre *site* e os arquivos do banco de dados.

Com a construção do banco de dados orais e sua implementação na rede *Internet* prevemos ainda a *interface* entre a sede do banco de dados em HTML[3] que dispõe de informações qualitativas, e os programas de análise acústica e estatística (PRAAT e MES), nos quais está sendo feito o trabalho de segmentação dos sinais.

A *interface* entre o *site* e esses programas permite ao usuário conhecer o método de análise acústica anterior a implementação dos *corpus* orais num hospedeiro da *Internet*. A figura 2 mostra a interligação entre o *site* e os programas de análise acústica.

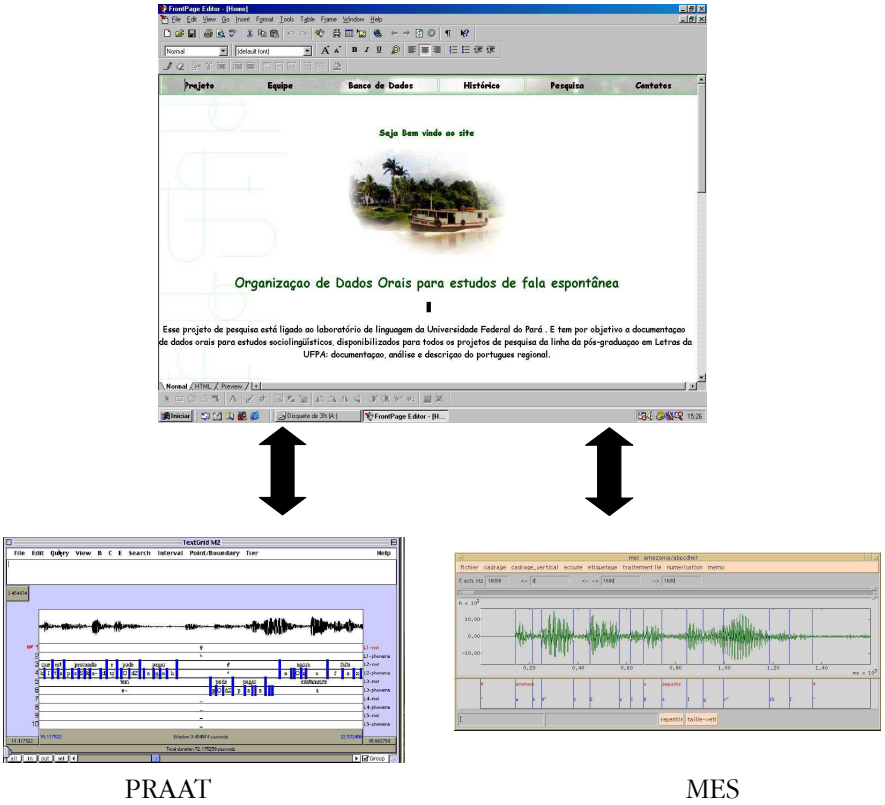


Figura 2 - interligação entre o *site* e o banco de dados e os programas de análise acústica.

Na fase atual, a aplicação e manutenção desse *site* na rede possibilita a divulgação desse trabalho destinado a fornecer informações lingüísticas para os estudiosos da linguagem e áreas afins. Seu caráter inovador se destaca por aplicar a *web* um banco de dados orais contendo fala espontânea, o que até então não poderia ser encontrado na rede, já que a grande maioria dos *sites* que tratam de *corpora* orais na *Internet* não conceberam tal empreendimento.

7. QUESTÕES TEÓRICAS

Diferentes escolas lingüísticas atestam uma importância ao status do *corpus* nos estudos lingüísticos (estruturalistas, gerativistas e sociolingüísticas). Por volta dos anos 50, somente Chomsky colocou em discussão a importância do uso de *corpora* como fonte primária de informação para as investigações lingüísticas. Todas as outras escolas atribuíram uma importância capital aos dados, apesar do fato de que elas não partilhavam da mesma concepção de *corpus* lingüístico em si (McENERY & WILSON, 1997).

A fonética, enquanto ramo da lingüística, apresenta este mesmo quadro. Sociolingüistas e foneticistas foram os primeiros a estudar a fala como objeto de estudos lingüísticos, particularmente quando eles se voltam para a fala espontânea. Assim, nós poderíamos estabelecer uma relação entre o vernáculo da sociolingüística e a fala espontânea para a fonética.

Entretanto, os estudos fonético-experimentais ainda se utilizam de fala lida de laboratório ou dados fonoestilísticos ou mesmo ainda dados produzidos em conversação controlada em laboratório em suas investigações. A causa maior é em decorrência da sensibilidade dos programas de análise que não são capazes de distinguir entre fala e os outros tipos de sons. Portanto, a qualidade das gravações compromete a naturalidade da fala. Eis a razão da utilização de gravações de alta fidelidade sem barulho anexo.

A acuidade fornecida pelas análises experimentais sobre os estudos dos sons da fala é inegável. Constitui-se um forte instrumento

sofisticado de precisão disponível para a pesquisa, quando nós precisamos tirar dúvidas que o ouvido humano não é capaz de solucionar. Este tipo de informação torna claro e dá conta de questões extralingüísticas indecifráveis por teorias fonológicas.

O problema maior do uso de dados espontâneos em trabalho experimental é a exigência de *corpus* de tamanho considerável, pois em sua totalidade são *corpus* com um controle bem menor do que os de fala artificial.

Entretanto, a fonética experimental deve ser contaminada pela concepção de língua da sociolingüística, e trabalhar muito mais com dados de língua em uso, em comunicação, em interação (LABOV, 1976).

A importância de banco de dados em tecnologia de fala é sentida em todos os níveis: (i) a disponibilidade de fonte de registro de fala devidamente documentado pode ser de grande utilidade para estudos diacrônicos futuros; (ii) a síntese de fala e o reconhecimento de voz tendem a utilizar muito mais sistemas <<data driven>>; (iii) o rumo epistemológico dos estudos sobre a linguagem, que priorizam muito mais conhecer a interelação entre os níveis do que os níveis em si, como dominou por muito tempo na lingüística e; (iv) o número cada vez maior de estudos interdisciplinares sobre a linguagem humana.

8. CONCLUSÃO

A organização e disponibilidade desse tipo de *corpus* é de extrema relevância para a comunidade científica. Poucos são aqueles projetos cujo *corpus* foi organizado em forma de base de dados orais e encontra-se disponível eletronicamente. E como se trata de um banco de dados orais destinado a estudos de variação lingüística, nada melhor do que documentá-lo devidamente e disponibilizá-lo para todos os projetos de pesquisa envolvendo o português falado. E já que, atualmente, são feitos investimentos pesados em *Internet,* acredita-se ser de suma importância a iniciativa de disponibilizar esses bancos de dados orais num *site* da rede *web.*

NOTAS

[1] Unidade sonora ininterrupta.

[2] http:/www.fbs.aust.com/aardvark.html. editor html com várias funções; http://www.infoflex.comau/flexed.htm. editor html; http://www.sede.com. sede *Internet* – serviços de hospedagem.

[3] http:/www.fbs.aust.com/aardvark.html. editor html com várias funções; http://www.infoflex.comau/flexed.htm. editor html; http://www.sede.com. sede *Internet* – serviços de hospedagem.

REFERÊNCIAS BIBLIOGRÁFICAS

CRUZ, R. Setting up spontaneous speech corpora. Comunicação oral apresentada *no Workshop Méthodes et Formalismes pour la Linguistique de Corpus*, realizado em Aix-en-Provence, Faculté des Lettres, Salle des Professeurs, no período de 12-13 octobre 2000a.

CRUZ, R. A*spects phonologiques et acoustiques du portugais parlé par des communautés noires de l'amazonie (Brésil)*. 2000b. Thèse (Doctorat) - Université de Provence.

CRUZ, R.; HIRST, D. Mise en oeuvre d'un corpus spontané. Trabalho apresentado em forma de painel no *RJC99 (Rencontre des Jeunes Chercheurs en Parole)*, Avignon (France), IUP Informatique d'Avignon, 18-19 de novembro, 1999.

CRUZ, R.; HIRST, D.; BEL, B. Mise en oeuvre d'un corpus spontané. *Revue Parole*. DUEZ, Danielle (Ed.). volume especial sur la parole spontanée.

LABOV, W. *Sociolinguistic*. Traduction d'Alain Kihm. Les Éditions de Minuit: Paris, 1976.

McENERY, T.; WILSON, A. *Corpus linguistic*. Edimburg University Press: Edinburg, 1997.

OLIVEIRA, M. *Prosodic Features in spontaneous narratives*. 2000. 241f. Dissertation (PhD) - Simon Frase University, 2000.

TRINDADE, R. *O som da fala dos pescadores de Cametá*. Memóire de Master du Département de Linguistic de l'Université Fédérale du Santa Catarina (Brésil).