ORGANIZAÇÃO DE CORPORA SOCIOLINGÜÍSTICO EM BANCO DE DADOS ORAIS PARA ESTUDOS DO PORTUGUÊS REGIONAL PARAENSE

Regina Célia Fernandes Cruz - UFPA
Helane de Fátima Gomes Fernandes - UFPA
Jailma do Socorro Uchôa Bulhões - UFPA
Léa da Silva Fernandes - UFPA

RESUMO

A organização de corpora em banco de dados orais é de interesse de todas as áreas que desenvolvem pesquisas sobre linguagem humana. Este trabalho apresenta as fases necessárias para a construção de banco de dados orais: (i) organização, (ii) segmentação fonética e (iii) o armazenamento de corpora sociolingüísticos.

PALAVRAS-CHAVE: Variação lingüística; fonética; banco de dados orais.

ABSTRACT

The organization of corpora in oral database interests all of the areas which develop researches about language. This essay traces considerations about the necessary stages to the constructions of a oral database: (i) the digitalization, (ii) phonetic segmentation and (iii) the storage of the linguistic corpora.

KEY WORDS: Linguistic variation; Phonetic; speech database.

INTRODUÇÃO

O objetivo principal do presente estudo é o de demonstrar a importância de organização de corpora orais em banco de dados e de como tal empreendimento facilita a 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, Cruz et alii 2002a e 2002b, e Cruz e Bulhões 2003). 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.

Ainda ressentem-se da existência de corpus organizado em forma de base de dados orais e disponível eletronicamente. Por essa razão, há necessidade de documentar e disponibilizar dados de fala espontânea representativos do português brasileiro a fim de atender os projetos de pesquisa da linha de pós-graduação em Letras de Documentação, Descrição e Análise do Português e servir de suporte para os estudos de reconhecimento de voz e de locutor desenvolvidos na área de Engenharia Elétrica.

Apesar de haver atualmente um crescimento no interesse dos estudos fonéticos em relação à fala espontânea, o fato de a metodologia da fonética experimental ainda vigente basear-se em corpus lido, associado às dificuldades de sensibilidade dos equipamentos utilizados nesse tipo de estudo, impossibilita sua aplicação imediata e direta a dados espontâneos. Este trabalho relata os procedimentos representados pela tentativa de facilitar o acesso automático a dados do português espontâneo em estudos na área da fonética experimental e engenharia da fala, através da organização de um corpus em banco de dados orais, contendo informações qualitativas e quantitativas dos dados. Quais são então os corpora em questão?

1 DESCRIÇÃO DO CORPORA EM QUESTÃO

O referido corpus constitui mais de 30 horas de gravação com amostras de três variedades lingüísticas do português: (i) o Português Regional Paraense (AM), coletado por Trindade (1992) no período de 1989 –1990 e o corpus coletado por Rodrigues (2003) ambos coletados na cidade de Cametá[1]; (ii) o Português Afro-brasileiro (ABP), variedade alvo, falada pelas comunidades quilombolas do Pará, coletado entre 1992-1996 por Cruz (2000b); (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 falada nas grandes capitais brasileiras, o corpus cedido por Neto (1998), coletado no período de agosto a novembro de 1998 em Belém e que compunha o acervo do projeto NURC-PA, sendo também composto de falantes de nível superior; e por último o de Cassique, coletado no ano de 2001, formado com amostras da variedade lingüística de falantes analfabetos[2].

[1] Este servirá de base para sua dissertação de mestrado sobre os aspectos de variação contidos no fonema /r/.

[2] Estando em processo de digitalização.

Dois aspectos os aproximam:

a) a língua: todos contém amostras do português brasileiro;
b) a metodologia seguida para sua formação: Os seis corpora foram formados prioritariamente para servir de suporte empírico à investigações sociolingüísticas.

Inicialmente, todos os dados sonoros que constituem o corpus desse 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, devido às más condições das fitas cassete, Cruz (2000b) procedeu a um inventário das cassete 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 (nota 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 de gravação e variedade lingüística. Todas essas informações figuram no código atribuído a cada sinal sonoro no momento da digitalização.

2 FASES DO PROCESSO

Como se trata de um corpus coletado em trabalho de campo, antes da concepção e posterior execução do banco de dados, foi feito um inventário do material lingüístico presente no mesmo. 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.

Este inventário foi fundamental para a organização dos corpora no banco de dados orais concebido e executado por Cruz (2000a, 2000b, Cruz et alii 2002, Cruz & Bulhões 2003).

Além do inventário feito, o corpus foi submetido a mais três outras fases de tratamento antes de figurar no banco dados orais. Estas fases são digitalização, segmentação e armazenamento, as quais passaremos a descrever.

2.1 DIGITALIZAÇÃO

O trabalho de digitalização possibilita a documentação do material lingüístico de forma segura e proporciona o acesso a este de forma rápida e eficaz, sendo um recurso viável para documentar material de fala humana: entrevistas, depoimentos etc., podendo ser utilizado em qualquer área de conhecimento que utiliza material desta natureza[3]. Esse procedimento torna-se uma alternativa ao armazenamento do material lingüístico em fitas de áudio, já que estas apresentam restrições em relação a sua conservação como: a vulnerabilidade das fitas de áudio a fatores climáticos, como a umidade, e as diferenças de rotação que tal material apresenta após permanecer muito tempo armazenado; outro obstáculo a esse tipo de documentação é o grande número de fitas acumuladas durante a pesquisa, assim se esse corpus for digitalizado, as fitas podem ser excluídas.

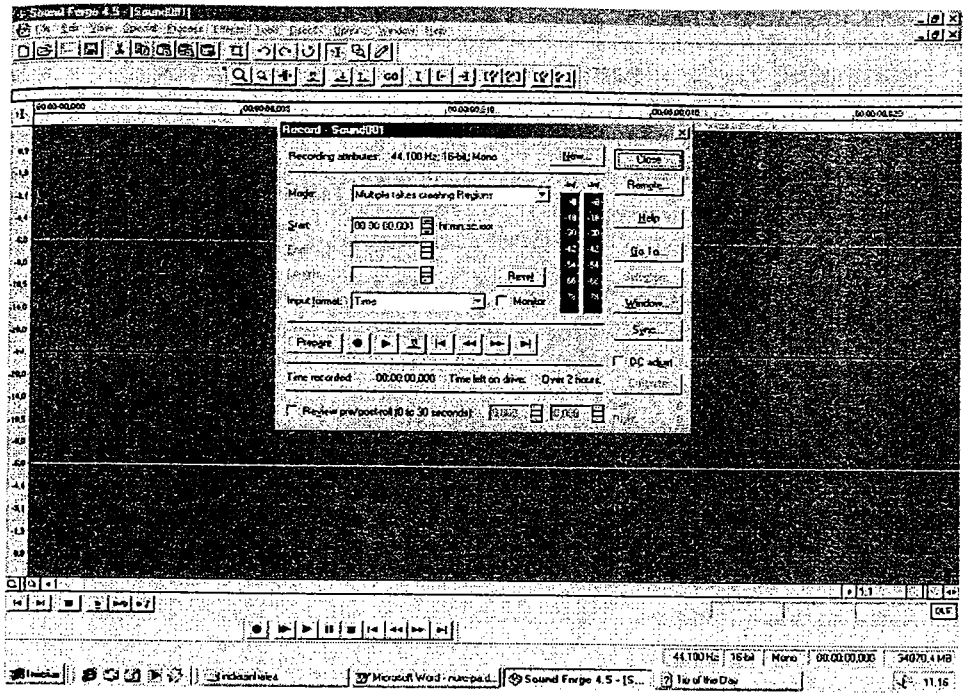
O SOUNDFORGE é um programa de tratamento de sons que possibilita a digitalização, com ele é possível manipular o controle de canais de volume com um equalizador que indica as possíveis saturações do som a ser gravado, possibilitando controle na altura da gravação a ser digitalizada, imprimindo qualidade sonora satisfatória a este material.

Este programa permitiu reunir duas ou mais tomadas de gravação que pertenciam a mesma situação de gravação e que por razões técnicas (limitação física do tempo da fita cassete por exemplo) se encontrava 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

[3] Pensa-se em particular em áreas como a Psicologia, Antropologia e Sociologia que também utilizam a linguagem verbal no estudo de seu objeto de pesquisa.

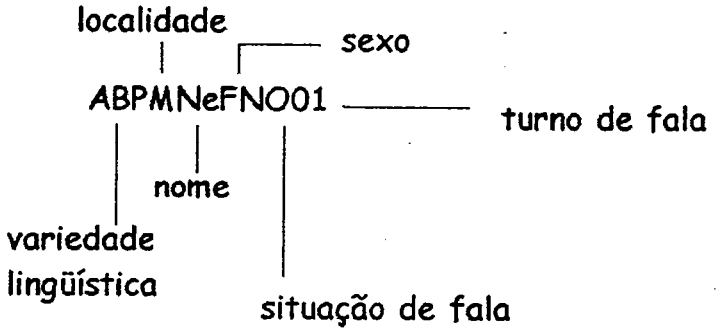
reunidas em um mesmo sinal e separadas por um intervalo de silêncio de 150 milisegundos. Foi igualmente possível eliminar barulhos indesejáveis para análise acústica.

Tela inicial do programa SOUNDFORGE,
onde se vê o equalizador para controle da qualidade sonora da gravação,
além dos comandos de CD Player.

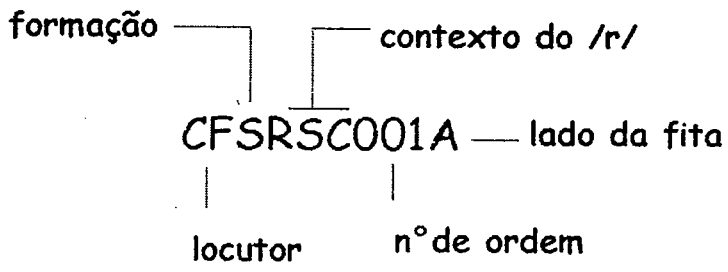


Para não se perder a traça das gravações originais, cada sinal gravado contém as seguintes informações: um número de ordem, a data da gravação e a duração total do sinal de áudio que ele contém, além de detalhes de cada arquivo digitalizado por SOUNDFORGE (assinalando data da digitalização, as tomadas 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: o nome do locutor, a localidade, a situação de fala, sexo do locutor e o turno de fala, seguindo a notação sugerida por Cruz (2000b) e Rodrigues (2003).

1 - Padrão de notação para o corpus armazenado
no banco de dados do projeto
"Vozes da Amazônia" (Cruz, 2000)



2 - Padrão de notação para o corpus coletado por Rodrigues (2003),
estipulado de acordo com os objetivos de sua análise
sobre a variação do /r/ na cidade de Cametá.



O processo de documentação exige duas fases distintas de tratamento do material gravado: o inventário das fitas de áudio, que consiste na audição das fitas cassete confirmando todas as informações contidas em suas respectivas capas, já que em viagens de campo, problemas como troca e mesmo extravio de fitas são passíveis de acontecer, tal recurso possibilitou uma pré-seleção das fitas que se enquadram melhor aos objetivos do projeto, no caso, gravações de fala espontânea, identificando a qualidade sonora destas e descartando várias fitas com qualidade ruim, e a digitalização, que é feita quando interliga-

se um aparelho de som stereo ao computador, e aciona-se os comandos RECORD CONTROL e PLAYCONTROL para a manipulação dos sons contidos nas fitas. Como se dá essa manipulação? Através do comando RECORD CONTROL que nada mais é que o equalizador do Sound Forge, controlando assim os recursos de gravação e possibilitando que o som seja colocado num volume em que não haja risco de saturação, ou seja, que ele não fique alto demais impedindo sua segmentação no programa PRAAT.

Quanto ao comando Play Control ele é responsável pelo controle geral do volume e está diretamente ligado à reprodução dos dados digitalizados, através dele podemos acionar todos os acessórios necessários à reprodução bastando para isso selecionarmos os dispositivos : **Line-2, Wave/DirectSc e Microfone** e deslizando a barra de controle até alcançarmos a freqüência desejada.

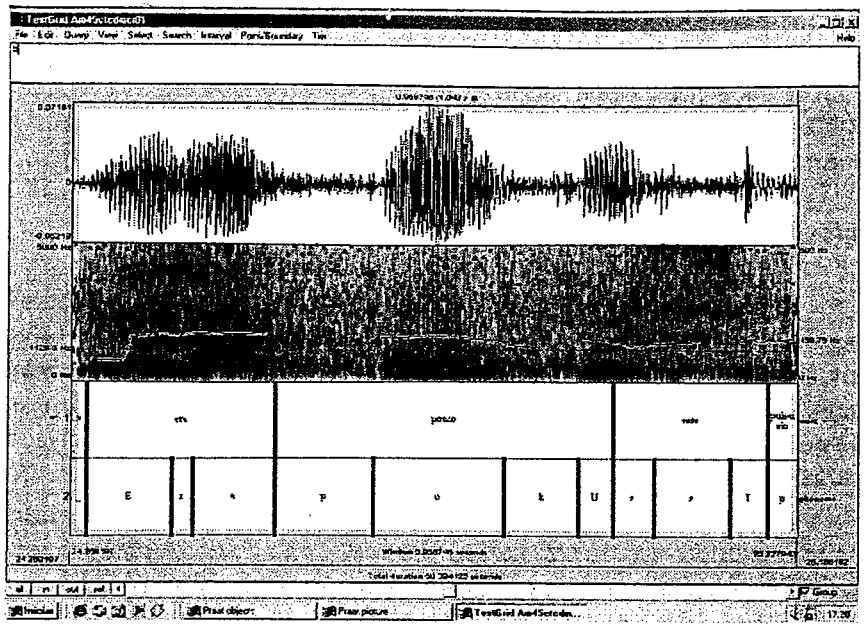
O programa SOUNDFORGE possui o mesmo caráter de editor do Microsoft Word, ou seja, as mesmas funções (colar, copiar, excluir, desfazer...) que propiciam no WORD a edição de um texto, podem ser feitas com o som, através deste programa. Após esse procedimento, o material sonoro está preparado para a próxima fase da formação do banco de dados: a segmentação.

2.2 SEGMENTAÇÃO

A segmentação é uma das etapas necessárias para que os dados já digitalizados sejam organizados de forma que fiquem prontos para serem utilizados em estudos de variação lingüística do português brasileiro, contendo informações tanto qualitativas quanto quantitativas dos sinais sonoros. Após a segmentação, o material lingüístico pode figurar em um banco de dados orais.

A tarefa de segmentação anteriormente era feita automaticamente utilizando o programa MBROLA, porém em decorrência de problemas causados justamente por esse tipo de segmentação é que desde 2001 recorre-se a uma segmentação feita manualmente.

Para a segmentação manual, atualmente utilizam-se dois instrumentos. O primeiro é o programa PRAAT, escolhido por possibilitar uma segmentação multilinear e dá informações quantitativas dos sinais sonoros.



Os PRAAT é um programa que pode ser obtido gratuitamente através do endereço eletrônico *http://fon.hum.uva.nl/praat/praat5133.html ou paul.boersmahum.uva.nl.* A figura 3.2.1 exemplifica uma segmentação manual no mesmo.

O segundo instrumento utilizado para a tarefa de segmentação manual é o alfabeto SAMPA. A opção por este alfabeto se deu pelo mesmo ser próprio para transcrição fonética em meio eletrônico, impedindo que ocorram problemas na transferência e na identificação dos símbolos em outras máquinas ou em diferentes meios de divulgação.

Todo o trabalho de segmentação de um sinal sonoro se inicia com a identificação dos turnos de fala e prossegue com a primeira segmentação diretamente no sinal sonoro e no espectograma que é a de retirada dos mesmos existentes em um sinal sonoro. Extraídos o turno de fala escolhido por sua qualidade que propicia a sua configuração em um banco de dados codifica-se o mesmo de forma que informe variedade lingüística, localidade, nome e sexo do seu informante, e ainda a situação de fala presente e o número do turno extraído. Após isso, passa-se para duas outras segmentações: a em palavras e a em fonemas.

No arquivo texto citado anteriormente encontra-se um nível para a segmentação em palavras e um outro nível para a segmentação em fonemas. Ao se identificar palavras se aponta seus limites com etiquetas diretamente no sinal sonoro. Após a delimitação de palavras, passa-se para a delimitação dos fonemas e o alinhamento de cada um ao seu correspondente no sinal sonoro. Por fim, é feita a revisão de cada segmentação e do alinhamento de fonemas.

O banco de dados orais de Cruz (2000b, Cruz et alii 2002a e 2002b, e Cruz e Bulhões 2003) contém 66 sinais sonoros. Antes da segmentação manual havia um total de 7 sinais sonoros fragmentados em turnos de fala e apenas 35 fragmentos segmentado fonética e ortograficamente. Atualmente, dos 66 sinais sonoros que formam o corpus de Cruz (idem) há 35 segmentados em turnos de fala e 80 fragmentos segmentados fonética e ortograficamente.

A segmentação dos sinais é uma atividade onerosa em tempo, mas necessária e primeira para levar-se a termo qualquer análise lingüística. E as informações quantitativa dos sinais sonoros armazenados no banco de dados orais do projeto são fornecidas pela tarefa de segmentação aqui descrita.

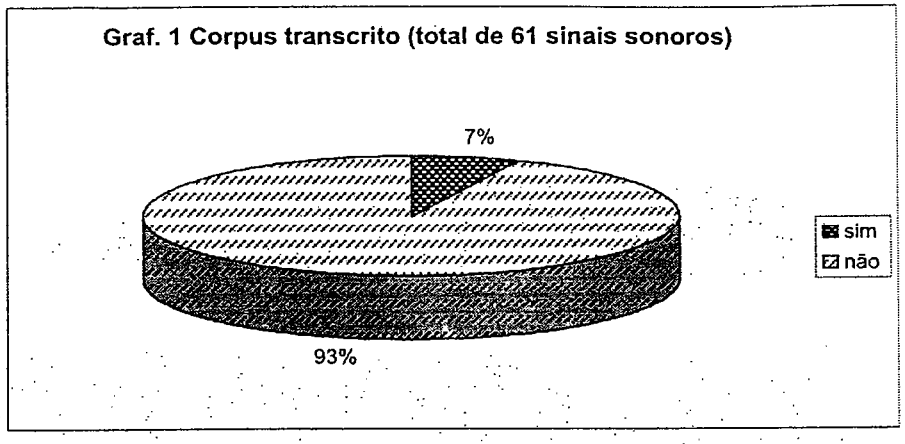
2.3 ARMAZENAMENTO

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:

2.3.1 Sinais Sonoros

Este arquivo contém informações diretas sobre os sinais sonoros, que correspondem a identificação do locutor, a duração do sinal sonoro, a situação de fala, a data, ao lugar e a 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. Este sinais sonoros correspondem a unidade tomada de gravação descrita no item 2.

O gráfico 3.3.1 mostra a porcentagem de sinais sonoros armazenados já transcrito em turnos de fala.



na figura 3.1.1. O arquivo de sinais sonoros contém as seguintes informações:

2.3.2 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, no Pará. No banco de dados há referência tanto à localidade de origem dos locutores quanto às localidades onde foram feitas as gravações.

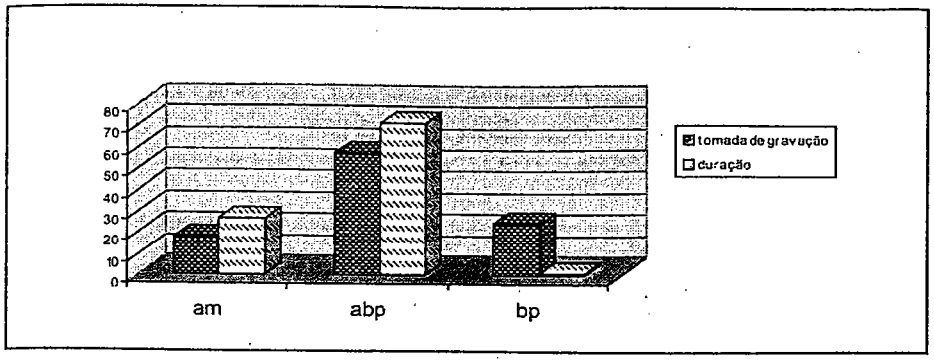
2.3.3 Variedades Lingüísticas

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

O gráfico 3.3.2 apresenta como as variedades estão distribuídas no banco de dados em número e duração de sinais sonoros.

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. 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. Os dados do português brasileiro padrão foram gravados em 1998 com falantes do português residentes em Vancouver no Canadá, 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).

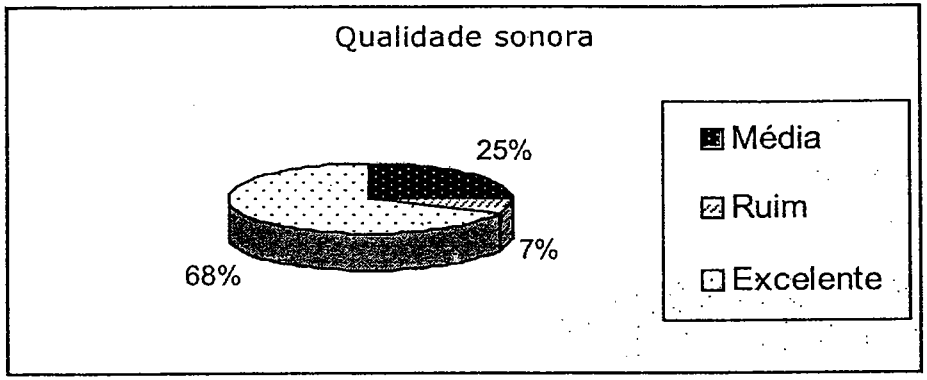
2.3.4 Qualidade Sonora

Descreve 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 que 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, obteve-se tomadas de gravação passíveis de serem utilizadas nesse tipo de estudo. Logo, essa

proporção deve aumentar em coleta de dados, visando tanto obter fala espontânea quanto gravações com alta fidelidade.



2.3.5 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. É com eles também que é possível serem realizadas as análises prosódicas de ritmo e entoação, pois sua unidade de base compreende um turno de fala, o qual consiste num enunciado de um único locutor.

Eles estão sendo transcritos com PRAAT para a parte segmental e a codificação MOMEL para a normalização da curva de Fo.

A transcrição segmental é feita com o alfabeto SAMPA.

Até o presente momento possuímos a segmentação de 80 fragmentos sonoros. Este arquivo contém os seguinte sub-arquivos:

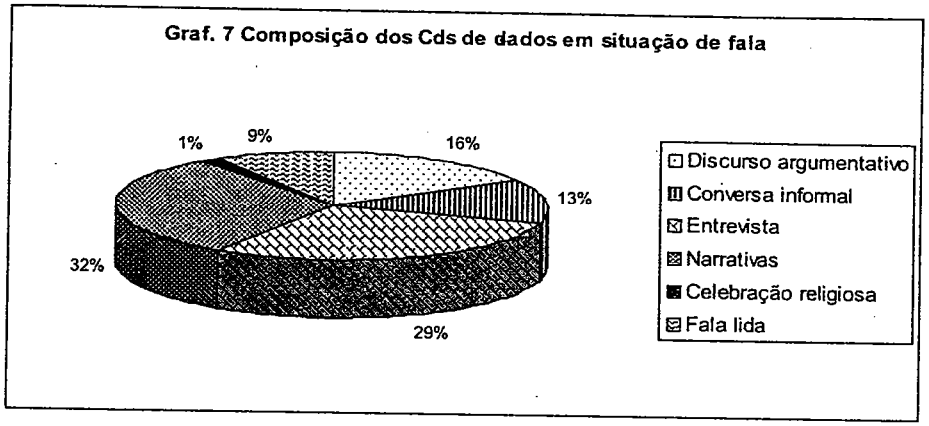
2.3.6 Locutores

Este arquivo apresenta informações sobre sexo, faixa etária, origem, escolaridade e variedade lingüística dos locutores. No corpus há um número maior de vozes masculinas. E há também uma maior representatividade de locutores na faixa de 30-70 anos de idade.

2.3.7 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.

Nós escolhemos, em um primeiro momento, 61 sinais de áudio, distribuídos da seguinte forma como expresso no gráfico 7:



3 QUESTÃO TEÓRICA

Diferentes escolas lingüísticas atestam uma importância ao status do corpus lingüístico. 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.

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 corpus nos estudos lingüísticos (estruturalistas, gerativistas e sociolingüísticas). 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 de sermos obrigados a utilizar gravações de alta fidelidade sem barulho anexo.

CONCLUSÃO

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 inter-relação 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.

A organização e disponibilidade desses 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 da linha de documentação, descrição e análise do português. E já que vivemos num tempo de investimentos pesados em Internet foi de suma importância a iniciativa de disponibilizar esses banco de dados orais num site da rede web.

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 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.

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

REFERÊNCIAS

CRUZ, Regina & BULHÕES, Jailma, 2003. *Implementação de um banco de dados orais destinado ao estudo do português regional paraense.* In: VII Encontro IFNOPAP: navegando entre o rio e a floresta. Tema: "Populações e tradições às margens do rio Tocantins: um diálogo entre a cultura e biodiversidade", Belém.

_____, R; Bulhões, J. & Fernandes, L., 2002b, "Banco de dados orais: uma nova perspectiva aos estudos sobre o português brasileiro", Comunicação oral apresentada no *I Congresso Internacional de Fonética e Fonologia / VII Congresso Nacional de Fonética e Fonologia*, realizado nos dia 28-30/10/2002, em Belo Horizonte (MG).

_____, R; Bulhões, J. & Fernandes, L., 2002a, "Organização de banco de dados orais para estudos de fala espontânea", Comunicação oral apresentada no *XVII Encontro Nacional da ANPOLL*, realizado nos dia 24-28/06/2002, em Gramado (RS).

_____ R. 2000a, "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 2000.

_____, R., 2000b, *Aspects phonologiques et acoustique du portugais parlé par des communautés noires de l'amazonie (brésil)*, thèse de doctorat, université de provence.

_____, R., & hirst, D., 1999, "Mise en oeuvre d'un corpus spontané", trabalho apresentado em forma de painel no *rjc99 (rencontre des jeunes chercheurs en parole)*, avignon (França), iup informatique d'avignon, 18-19 de novembro.

_____, R., hirst, D. & bel, B., "Mise en oeuvre d'un corpus spontané", in *revue parole*, volume especial sur la parole spontanée, danielle duez (ed.).

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

McENERY, T. & WILSON, A., 1997. *Corpus Linguistic*. Edimburg University press: Edinburg.

OLIVEIRA, M. 2000. *Prosodic Features in Spontaneous Narrativas*, PhD. Dissertation, Simon Frase University, 241 p.

TRINDADE, R., 1992. *O Som da fala dos pescadores de Cametá*. Memóire de Master du Département de Linguistic de l'Université Fédérale du Santa Catarina (Brésil).

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.

http://fon.hum.uva.nl/praat/praat5133.htm