Formação e Anotação do *Corpus* do Projeto AMPER Norte

Regina CRUZ (UFPA/CNPq)
Ilma Pinto do Espírito Santo (Aluna de Mestrado CML/UFPA)
Camila Brito (Bolsista PIBIC/CNPq)
Rosinele Lemos (SEDUC-PA)
Isabel Cristina Rocha dos Remédios (Aluna de Mestrado CML/UFPA)
João Freitas (Aluno de Mestrado CML/UFPA)
Elizeth Guimarães (Aluna de Mestrado CML/UFPA)

RESUMO

O presente trabalho tem como objetivo principal apresentar como estão sendo organizados, tratados e anotados os *corpora* formados para o estudo das características prosódicas das variedades linguísticas do português falado na Amazônia Paraense. Trata-se de um estudo vinculado ao Projeto AMPER que tem como objetivo principal fornecer a caracterização acústica e prosódica das línguas românicas, assim como um atlas multimédia on-line (CONTINI *et al* 2002: 227-230; MOUTINHO *et al* 2001: 245-252). No que diz respeito ao português, quinze instituições participam na descrição das suas três principais variedades, a saber: o português europeu continental, o português europeu insular e o português brasileiro.

A UFPA já participa do referido projeto, desde de 2007, tendo sob sua responsabilidade a confecção do Atlas Prosódico Multimédia da Região Norte do Brasil. Atualmente, cinco atlas estão em andamento: a) de Belém (BRITO, em andamento; GUIMARÃES, em andamento); b) de Abaetetuba (REMÉDIOS, em andamento); c) do Marajó (FREITAS, em andamento), de Baião (LEMOS, em andamento) e de Cametá (SANTO, em andamento).

Como um dos objetivos do projeto AMPER compreende uma análise contrastiva dos dialetos estudados, o *corpus* gravado é formado de seis repetições de 66 frases do *corpus* de base do projeto para a língua portuguesa. Cada um dos elementos constituintes das frases possui uma imagem correspondente, uma vez que não é permitido nenhum contato dos informantes com as frases escritas. A série de frases que forma o *corpus* do projeto AMPER obedece a critérios fonéticos e sintáticos previamente estabelecidos.

Uma vez que nas vogais reside a maior parte da informação relevante no que concerne à curva prosódica e, tendo-se em conta as características da estrutura acentual do português, escolheram-se vocábulos representativos das diversas estruturas acentuais (oxítona, paroxítona e proparoxítona) nas diversas posições frásicas.

Sintaticamente as frases foram montadas de forma a apresentar Sujeito - Verbo – Complemento (SVC). Com relação a entoação, elas foram concebidas de modo a contemplar as modalidades declarativas e interrogativas globais. Portanto, as frases utilizadas nas gravações são do tipo SVC e suas expansões com a inclusão de Sintagmas Preposicionais. Quanto à estrutura sintática, todas as frases possuem apenas: 1) três personagens: Renato, pássaro e bisavô; 2) três sintagmas adjetivais: nadador, bêbado e pateta; 3) três sintagmas preposicionais indicadores de lugar: de Mônaco, de Veneza e de Salvador; e 4) um único verbo: gostar.

No momento da coleta de dados, a cada informante são pedidas seis repetições da série de frases do *corpus* (em ordem aleatória), sendo selecionadas para análise acústica as três melhores repetições, a fim de serem estabelecidas médias dos diversos parâmetros acústicos: duração, frequência fundamental (F0) e intensidade.

Conforme determina o projeto geral, para a seleção dos informantes, levam-se em consideração os seguintes critérios: 1) faixa etária (acima de 30 anos); 2) escolaridade (fundamental, médio e superior); e 3) tempo de residência na localidade (nativos do local). A partir desses critérios, são selecionados seis informantes, três homens e três mulheres, que participaram da coleta de dados. Trata-se, portanto, de uma amostra estratificada. Cada informante recebe um código, que contem informações sobre seu perfil.

Ao todo obtivemos seis sinais sonoros por variedade investigada. A taxa de amostragem de cada sinal é de 44.100 Hz, 16 bits, sinal mono. Uma vez a gravação concluída, procede-se à separação por informante das 396 frases do sinal original em um arquivo sonoro específico.

O material gravado sofre cinco etapas de tratamento: a) codificação das repetições; b) segmentação vocálica dos sinais selecionados no programa PRAAT 5.0; c) aplicação do *script* praat; d) seleção das três melhores repetições e; e) aplicação da interface Matlab para se obter as médias dos parâmetros das três melhores repetições.

No caso da codificação das repetições, retomou-se o código do informante, contendo o seu perfil, acrescentou-se o código de cada frase já estabelecido pelo projeto AMPER, contendo as indicações sintáticas, fonéticas e prosódicas, por último acrescenta-se um número de ordem cronológica da repetição.

Para o trabalho de segmentação fonética, utilizamos o programa PRAAT. Apenas um nível de segmentação fonética é criado, denominado de <vogais>, uma vez que o projeto AMPER determina indicar apenas as realizações vocálicas de cada sinal sonoro analisado. Da mesma forma o *script* automático criado para o PRAAT pela coordenação do AMPER considera apenas as realizações vocálicas para efetuar os cálculos dos parâmetros acústicos controlados para análise. O *script* PRAAT criado para análise acústica lê como códigos apenas a letra “v” (indica vogais plenas) e a letra “f” (indica vogais fracas ou elididas). Durante a segmentação fonética, estabelecemos as

escalas de *pitch* adequadas para a análise de cada informante. Para os falantes do sexo masculino esta escala fica entre 50 Hz a 250 Hz e para os falantes do sexo feminino de 110 Hz a 370 Hz.

Uma vez concluída a segmentação fonética dos 396 sinais sonoros de cada informante, passa-se à aplicação do *script praat* criado para o projeto AMPER. O *script praat* foi aplicado a cada uma das 396 frases por informante. A aplicação desse *script* gera um arquivo .TXT contendo as medidas dos parâmetros acústicos (intensidade, frequência fundamental, intensidade e formantes) das vogais de cada repetição.

Antes de se proceder a análise acústica na interface Matlab, selecionam-se as três melhores repetições de cada frase em termos de qualidade sonora e de similaridade de distribuição de vogais plenas (v) e elididas (f).

A aplicação da interface Matlab fornece a média dos parâmetros físicos – F0, duração e intensidade – em um arquivo fono.txt das três repetições de cada frase e das duas modalidades. A interface gera mais outros arquivos em formato de imagem contendo gráficos das médias de F0, duração e intensidade de cada modalidade individualmente, assim como gráficos comparativos de ambas as modalidade. A interface gera igualmente arquivos .ton contendo uma síntese de cada modalidade sem a parte segmental.

Portanto, o *corpus* do projeto AMPER Norte é composto de **198** frases, totalizando **1.188** frases por informante, contendo amostras das variedades linguísticas faladas em Belém, Cametá, Abaetetuba, Baião e de Curralinho.

Referências Bibliográficas

BRITO, Camila. *Atlas prosódico multimédia do Português do Norte do Brasil – AMPER-POR: variedade lingüística da zona rural de Belém (PA)***,** Belém**:** UFPA/ILC/FALE, **em andamento** (Plano de Iniciação Científica).

CONTINI, Michel *et al*. "Un Projet d'Atlas Multimédia Prosodique de l'Espace Roman". In: Bel, B. & Marlien, I. (edd.): *Proceedings of the 1st International Conference on Speech Prosody*. Aix-en-Provence: Laboratoire Parole et Langage, **2002**, 227-230.

FREITAS, João. *Atlas Prosódico Multimédia do Município da ilha do Marajó (PA)*, Belém: UFPA/ILC/CML, em andamento (Dissertação de Mestrado).

GUIMARÃES, Elizeth. *Atlas Prosódico Multimédia da Belém Insular (PA)*, Belém: UFPA/ILC/CML, em andamento (Dissertação de Mestrado).

LEMOS, Rosinele. *Atlas Prosódico Multimédia do Município de Baião (PA)*, Belém: UFPA/ILC, em andamento (Projeto Dissertação de Mestrado).

MOUTINHO, Lurdes de Castro *et al*. Contribuição para o estudo da variação prosódica do Português Europeu. In: Sánchez Miret, F. (ed.): Actas do XXIII CILFR (Salamanca, Espagna, 22-28 Set. 2001). Vol. 1. Tübingen: Niemeyer, 2001, 245-252.

REMÉDIOS, Isabel. *Atlas Prosódico Multimédia do Município de Abaetetuba (PA)*, Belém: UFPA/ILC/CML, em andamento (Dissertação de Mestrado).

SANTO, Ilma. *Atlas Prosódico Multimédia do Município de Cametá (PA)*, Belém: UFPA/ILC/CML, em andamento (Dissertação de Mestrado).

Palavras-chave: *corpus* oral, entoação modal, projeto AMPER, português brasileiro.

Trabalho para Comunicação Oral