CRITÉRIOS LINGÜÍSTICOS PARA A ELABORAÇÃO DE LISTAS DE PALAVRAS NO PORTUGUÊS BRASILEIRO
(LINGUISTIC CRITERIA FOR THE DEVELOPMENT OF WORD LISTS IN BRAZILIAN PORTUGUESE)

Aglael GAMA-ROSSI (Pós-Doutoranda-PUCSP/FAPESP) & Adelaide SILVA (DELIN/UFPR; LAFAPE/UNICAMP)

ABSTRACT: *The paper discusses linguistic criteria underlying the development of word lists in Brazilian Portuguese. It presents the results of a survey about the frequency of grammatical categories, word lengths, stress and syllable patterns, using the LISTAS program, which is based on the MiniDicionário Aurélio*.

KEYWORDS: *linguistic criteria; word lists; Brazilian Portuguese.*

0. Introdução

Nos últimos anos, o Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE-IEL/UNICAMP) tem sido procurado por audiologistas que buscam suporte para o balanceamento fonético de listas de palavras e sentenças no português brasileiro (PB), utilizadas na logoaudiometria[1]. Graças ao trabalho de Albano, Moreira, Silva, Aquino e Kakinohana (1999 [1995]), foi possível um novo entendimento acerca do balanceamento fonético, antes restrito à presença de todos os fonemas da língua na lista a ser balanceada. Os autores desenvolveram um *software*, o ORTOFON, o qual permite que textos digitados ortograficamente sejam convertidos numa transcrição fonética larga e tenham seus fonemas contados. Usando o ORTOFON, eles realizaram um levantamento da freqüência de ocorrência dos fonemas no PB, com base em dois *corpora*, o *MiniDicionário Aurélio* e um conjunto de 57 gravações do Projeto NURC. O balanceamento fonético proposto no LAFAPE passou então a ser feito de acordo com os seguintes passos: a lista a ser balanceada passa pelo ORTOFON, onde é transcrita foneticamente e tem seus fonemas contados, e após isso é comparada ao *corpus* de referência (*MiniDicionário Aurélio*, para listas de palavras, e NURC, para listas de sentenças), por meio de uma estatística não-paramétrica, o teste de *Spearman*. Assim, considera-se que a lista está foneticamente balanceada quando é obtida uma correlação próxima a 100% entre a distribuição da freqüência de ocorrência dos fonemas na lista e a distribuição da freqüência de ocorrência dos fonemas na língua (*corpus* de referência).

[1] De acordo com Zubick, Irizarry, Rosen, Feudo, Kelly e Strome (1983:88-89), a logoaudiometria refere-se aos procedimentos clínicos usados para mensurar dois tipos de estímulos de fala: o *SRT* (*Speech ReceptionThreshold*, cuja sigla é mantida em português) avalia a autenticidade das respostas previamente obtidas para tons puros, dentro do espectro de fala (500, 1000 e 2000Hz), registrando o ponto em que 50% das palavras selecionadas podem ser repetidas, e a Discriminação de Palavras (*Word Discrimination*) fornece a percentagem de palavras corretamente percebidas, geralmente apresentadas entre 25 e 40 dB NS.

Entretanto, o contato com as listas trazidas ao LAFAPE fez-nos levantar questões acerca da necessidade de controlar outros critérios além do balanceamento fonético, o que foi confirmado por um levantamento bibliográfico sobre os critérios adotados na elaboração de listas de palavras empregadas na logoaudiometria de outras línguas. Assim, neste trabalho, o objetivo é dar um primeiro passo na elaboração de listas de palavras no PB. Para isso, serão apresentados os principais resultados de um levantamento de freqüência de ocorrência usando o programa LISTAS, desenvolvido no LAFAPE, o qual permite fazer o levantamento da freqüência de ocorrência das palavras do *MiniDicionário Aurélio* em função da classe gramatical, do número de sílabas e dos padrões acentual e silábico.

1. Breve consideração sobre critérios adotados em listas de palavras em outras línguas

A literatura sobre a elaboração de listas para a logoaudiometria revela que os estudos partem dos critérios que serviram de base às primeiras listas elaboradas nas línguas indo-européias, designados por *critérios do modelo indo-europeu* (Nsamba, 1979), os quais dominaram a logoaudiometria mundial por muito tempo, para, logo em seguida, discutir e abandonar tais critérios em detrimento de listas que reflitam as características de suas línguas, e que, por isso, sejam adequadas para testar suas populações. Nsamba (op. cit.), por exemplo, coloca que, diferentemente do inglês, nas línguas Bantu, os monossílabos ocorrem em pequena quantidade e, além disso, quando são formados por C+semivogal+V (p.e., "kwe"), sua pronúncia isolada, em listas de palavras, não soa de modo natural, o que pode comprometer os *scores* de discriminação obtidos na logoaudiometria, principalmente em indivíduos com perda auditiva. Da mesma forma, embora haja uma grande quantidade de dissílabos nas línguas Bantu, é difícil, segundo ele, encontrar dissílabos que se aproximem dos espondeus das línguas indo-européias, com duas sílabas igualmente acentuadas. Por outro lado, as línguas Bantu apresentam uma grande quantidade de palavras *quasi*-monossilábicas[2], e, para Nsamba (op. cit.), utilizar tais palavras é necessário para o desenvolvimento de uma logoaudiometria mais representativa para os falantes de tais línguas.

É interessante notar que outros fatores podem interagir com critérios lingüísticos na seleção das palavras das listas utilizadas na logoaudiometria. Ashoor e Prochazka (1982) mantiveram, no árabe moderno padrão, um critério fundamental no inglês, o de listas de palavras formadas por monossílabos. Entretanto, o argumento para isso, no caso do árabe, parece ter sido um fator psico-lingüístico, e não a alta freqüência de ocorrência de monossílabos na língua, uma vez que os autores afirmam que a opção por monossílabos baseia-se no fato de que detectar palavras monossilábicas é mais fácil do que detectar palavras polissilábicas, porque as primeiras requerem menos tempo para serem percebidas. Um fator sócio-lingüístico, palavras igualmente familiares entre as classes instruídas e não-instruídas, foi também adotado para que as listas pudessem ser

[2] O autor refere que as palavras *quasi*-monossilábicas podem ser de dois tipos, a saber: CCV (p.e., "ggi") e N+CV (p.e., "nze").

usadas na avaliação logoaudiométrica da população árabe saudita como um todo. As palavras que compunham as listas foram retiradas de livros escolares, jornais e livros infantis, sendo evitados termos técnicos e científicos, e, sempre que possível, elas deveriam ter formas similares no árabe coloquial e no árabe moderno padrão. Elas deveriam ainda ser homogêneas quanto ao grau de dificuldade. No que se refere aos critérios lingüísticos, todas as palavras eram substantivos e as listas foram foneticamente balanceadas de acordo com a freqüência de ocorrência dos fonemas em *corpora* da língua.

Um outro exemplo de fatores psico e sócio-lingüísticos interagindo com critérios lingüísticos na construção de listas de palavras pode ser visto no trabalho de Mukari e Fiad (1991), os quais elaboraram listas de palavras em malaio, utilizando dissílabos, com estrutura silábica CVCV, muito freqüentes na língua. Para tornar as listas adequadas à avaliação logoaudiométrica da população malaia como um todo, os dissílabos CVCV foram inicialmente selecionados a partir de um dicionário, mas o principal critério para a construção das listas foi o uso deles na fala coloquial. Posteriormente, os dissílabos foram ainda submetidos a um teste de familiaridade, do qual participaram adultos de diferentes raças de nacionalidade malaia. As palavras tidas como pouco familiares foram eliminadas e o conjunto final foi subdividido em diversas listas, cada uma delas submetida a balanceamento fonético.

2. Critérios lingüísticos para a elaboração de listas de palavras no português brasileiro

Como um primeiro passo na elaboração de listas de palavras no PB, foi feito um levantamento da freqüência de ocorrência de classes gramaticais, número de sílabas e padrões acentual e silábico, utilizando-se o programa LISTAS que, conforme já foi mencionado, tem como base de dados o *MiniDicionário Aurélio*.

No que se refere às classes gramaticais, aproximadamente 98,5% das palavras do *MiniDicionário Aurélio* são palavras de conteúdo, subdivididas em: 58% de substantivos, 23% de adjetivos e 17,5% de verbos. Os outros 1,5% correspondem a palavras gramaticais, sendo que há 0,5% de advérbios e 1% de artigos, pronomes, preposições, conjunções, numerais e interjeições. Quanto ao número de sílabas, 94,5% do conjunto de substantivos, adjetivos e verbos variam de duas a cinco sílabas, sendo: 15% de dissílabos, 34,5% de trissílabos, 30% de quadrissílabos e 15% de pentassílabos. Os monossílabos correspondem a cerca de 0,6%, e as palavras de seis a dez sílabas somam os 4,9% restantes. Ao considerar esses dados, no que se refere à classe gramatical, a questão que surge é se as listas para a logoaudiometria no PB têm de ser compostas apenas da classe gramatical mais freqüente, ou seja, exclusivamente de substantivos. No que diz respeito à extensão vocabular, embora trissílabos e quadrissílabos ocorram em maior quantidade, coloca-se a questão de se é possível desconsiderar dissílabos (15%) e pentassílabos (15%). Portanto, os dados sobre classe gramatical e número de sílabas apontam para a discussão de se as listas de palavras no PB devem ser constituídas exclusivamente de elementos da classe gramatical e da

extensão vocabular mais freqüentes ou devem conter proporcionalmente elementos das principais classes gramaticais (substantivos, adjetivos e verbos) e extensões vocabulares (di, tri, quadri e pentassílabos).

Considerando-se o conjunto de palavras das três principais classes gramaticais, quanto à incidência do acento -antepenúltima, penúltima e última sílaba, temos: 11% de proparoxítonos, 54% de paroxítonos e 35% de oxítonos. Note que, no *MiniDicionário Aurélio*, os verbos são oxítonos, uma vez que são listados na forma infinitiva. Vejamos, então, o que ocorre com o total de oxítonos, paroxítonos e proparoxítonos, subdivididos em função das três classes gramaticais e das cinco extensões vocabulares mais freqüentes:

Tabela 1: Distribuição, em valores percentuais, das palavras oxítonas, paroxítonas e proparoxítonas em função das principais classes gramaticais e extensões vocabulares

Classe Gramatical	Número de Sílabas	Padrão de Acento (Valores Percentuais)		
		Oxítonos	Paroxítonos	Proparoxítonos
Substantivos	2	3	7	-
	3	5	13	2
	4	3	11	2,5
	5	2	5	2
Adjetivos	2	1	1	-
	3	1,5	4	1
	4	1	6	1,5
	5	1	3	1
Verbos	2	3	-	-
	3	8	-	-
	4	5	-	-
	5	1	-	-

Quando se considera a coluna dos paroxítonos, os dados sugerem uma resposta afirmativa às questões postas acima, de se uma lista de palavras no PB deve refletir proporcionalmente as principais classes gramaticais e extensões vocabulares. Assim, por exemplo, quanto à classe gramatical, embora os substantivos prevaleçam, é difícil desconsiderar os 6% de adjetivos quadrissílabos paroxítonos. Quanto à extensão vocabular, o mesmo ocorre para os 7% de substantivos dissílabos e os 5% de substantivos pentassílabos paroxítonos. No que concerne aos oxítonos, além dos verbos, predominantemente tri e quadrissílabos, aparecem em número não desprezível os substantivos trissílabos. Porém, quanto aos proparoxítonos, sua pulverização entre substantivos e adjetivos, tri, quadri e pentassílabos, sugere que eles não precisam comparecer nas listas de palavras do PB.

Por fim, um primeiro levantamento dos padrões silábicos mostrou que há um predomínio de sílabas do tipo CV, seguidas pelas sílabas de tipo CVC, V, VC e CCV. Sílabas pesadas, tais como: CCVC ou CCVCC, de um modo geral, parecem ser evitadas.

3. Considerações Finais

A nosso ver, os dados expostos acima são apenas um primeiro passo na elaboração de listas de palavras no PB. Pode-se perguntar se as ocorrências abaixo de 5%, na Tabela 1, devem ou não ser consideradas. A resposta a essa questão somente será dada a partir do emprego dos dados sobre classe gramatical, extensão vocabular, padrão de acento e padrão silábico no exame de listas de freqüência de ocorrência de palavras no PB. Além disso, dada a diversidade de classes sociais e as diferenças de escolaridade entre os indivíduos da sociedade brasileira, torna-se imprescindível que quaisquer listas de palavras elaboradas para testes clínicos sejam submetidas a testes de familiaridade, aplicados num conjunto de indivíduos que constitua uma amostra significativa dessas diferenças. Eventualmente, diferenças dialetais podem intervir no uso de uma determinada palavra de uma região para outra. Prevê-se ainda que listas elaboradas para testar adultos não sejam adequadas para testar crianças. Por fim, no que concerne ao balanceamento fonético, ele seria então o último passo na elaboração de tais listas, e seu uso deve ser redimensionado (Zubick, Irizarry, Rosen, Feudo, Kelly e Strome, 1983; Picard, 1984) frente não apenas a outros critérios lingüísticos, mas também a fatores sócio e psico-lingüísticos.

RESUMO: O trabalho discute critérios lingüísticos subjacentes à elaboração de listas de palavras no português brasileiro. São apresentados os resultados de um levantamento da freqüência de ocorrência das classes gramaticais, extensões vocabulares, padrões acentuais e silábicos, usando o programa LISTAS, que tem como banco de dados o *MiniDicionário Aurélio*.

PALAVRAS-CHAVE: critérios lingüísticos; listas de palavras; português brasileiro.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBANO, E. C.; MOREIRA, A.; SILVA, A.; AQUINO, P.; KAKINOHANA, R. Um conversor ortográfico-fônico e uma notação prosódica mínima para síntese de fala em língua portuguesa. In: E. SCARPA (org.). *Estudos de Prosódia no Brasil*. Campinas: Editora da UNICAMP, 1999 [1995].

ASHOOR, A. A.; PROCHAZKA, T. Saudi Arabic speech audiometry. *Audiology*, v.21, n.6, p.493-508, November/December, 1982.

MUKARI, S. Z. M.; SAID, H. The development of Malay speech audiometry. *The Medical Journal of Malaysia*, v.46, n.3, p.262-8, September, 1991.

NSAMBA, C. Luganda speech audiometry. *Audiology*, v.18, n.6, p.513-21, November/December, 1979.

PICARD, M. L'audiometrie vocale au Québec français. *Audiology*, v.23, n.4, p.337-65, July/August, 1984.

ZUBICK, H. H.; IRIZARRY, L. M.; ROSEN, L.; FEUDO, P.; KELLY, J.; STROME, M. Development of speech-audiometric materials for native Spanish-speaking adults. *Audiology*, v.22, n.1, p.88-102, January/February, 1983.