X ENCONTRO DE LINGUÍSTICA DE CORPUS
Belo Horizonte, 11 e 12 de novembro de 2011

Análise da frequência de ocorrência em pequenos corpora: uma proposta metodológica de comparação de corpora distintos *

Alan Jardel de Oliveira

Orientadora: Profª Drª Maria do Carmo Viegas

* O estudo é parte do Projeto Varfon-Minas (Variação Fonológica em Minas Gerais)

Contextualização

- Oliveira (2006): análise variacionista (cf. Labov, 1972) da sílaba final átona composta por consoante lateral alveolar mais vogal no falar de Itaúna.

- Formas variantes: a sílaba plena, o apagamento da vogal, a velarização da lateral após apagamento da vogal e o apagamento da sílaba.

- Objetivo: identificar e analisar os fatores que favorecem a realização das variantes. Um desses fatores pode ser a frequência de ocorrência da palavra.

Realização das variantes

1. Realização da plena da sílaba lV.
   Ex.: “['eli] (ele) não é bom profissional” (CH33).

2. Apagamento da vogal na sílaba lV.
   Ex.: “não arrepende não ... ['pel] (pelo) contrário” (LM17).

3. Velarização de /l/ ocorrida após o apagamento da vogal na sílaba lV.
   Ex.: “cortou o [ka'beł] (cabelo) ['deł] (dele).” (AH18).

4. Apagamento da sílaba lV.
   Ex.: “mas muitas vezes [e] (ele) tem que trabalhar junto com os alunos” (EM39)

Por que considerar a frequência de ocorrência das palavras?

- Bybee (2001):
  - A frequência de uso de palavras ou grupo de palavras em uma língua afeta a representação mental e a forma fonética das palavras ou grupo de palavras.
  - As mudanças sonoras foneticamente motivadas ocorrem primeiramente na palavras (ou grupos de palavras) de frequência de ocorrência mais alta do que nas de frequência mais baixa.

Objetivos do trabalho

- Apresentar uma proposta metodológica de comparação da frequência de ocorrência de palavras em corpora distintos.

- Averiguar se há correlação entre a frequência das palavras em corpora de fala do português brasileiro coletados em regiões distintas e entre corpora orais e escritos coletados em uma mesma região.

Desafios para mensuração da frequência de ocorrência

- Como contar a frequência de ocorrência das palavras em uma comunidade de fala?
  - Coleta e transcrição dos dados de fala espontânea.
  - O tamanho da amostra de dados de fala espontânea
    - 1 horas de entrevista transcrita corresponde a aproximadamente 5 mil palavras.
  - O assunto da entrevista pode tendenciar as frequências
    - Exemplo: a palavra *asilo* ocorreu 631, sendo que a média de ocorrência das palavras é 14 (a estrutura da entrevista levou os informantes a produzirem a palavra *asilo*)
  - Como saber se frequência observada em uma amostra corresponde à frequência na língua?

Análise da frequência baseada em outros corpora

- Pode-se considerar dados de fala de outras regiões?
  - Há grandes corpora de fala do PB já constituídos e disponíveis para consulta.
  - Há correspondência na frequência de ocorrência de uma palavra em comunidades distintas?

Análise da frequência baseada em outros corpora

- Pode-se considerar dados de escrita da mesma região?
  - A constituição do corpus de escrita é mais simples do que do corpus de fala.
  - Há correspondência na frequência de ocorrência de uma palavra em registros distintos?

Metodologia

- Corpora utilizados na análise
  - Corpus de fala espontânea coletado na cidade de Itaúna/MG
    - Entrevistas sociolinguísticas (cf. Labov, 1972)
    - 16 informantes (8 homens e 8 mulheres)
    - Aproximadamente 16 horas de gravação
    - 76.027 palavras (corpus pequeno, cf. Sardinha, 2000)
    - Estudo sobre a variação na sílaba final átona lV (cf. Oliveira (2006), Viegas e Oliveira (2008), Viegas e Oliveira (2009) e Oliveira (2011))

Metodologia

- Corpora utilizados na análise
  - Corpus escrito de Itaúna
    - Dados retirados de exemplares digitalizados dos 4 jornais da cidade
    - 1.997.942 palavras (corpus médio-grande, cf. Sardinha, 2000)
  - Corpus de fala do LAEL
    - Transcrição de dados de fala
    - 2.892.505 palavras (corpus médio-grande, cf. Sardinha, 2000)

Metodologia

- Contagem da frequência: *Wordsmith (Wordlist)*
- A análise estatística: *SPSS v.19*
- Método estatístico: *Análise de correlação* (mede o grau de associação linear entre duas variáveis contínuas)
- Estruturação do banco de dados comparativo
  - Seleção das palavras (terminadas em sílaba lV átona)
  - Criação da tabela

Recorte da tabela

PALAVRAS	JORNAIS	LAEL	FALA
ELE	3263	23104	1112
ESCOLA	901	1860	101
AULA	53	951	28
BAILE	91	725	2
SALA	456	444	31
CAVALO	39	436	1
NOVELA	21	430	8
BOLA	116	379	34
VALE	376	271	0
FALO	22	266	37

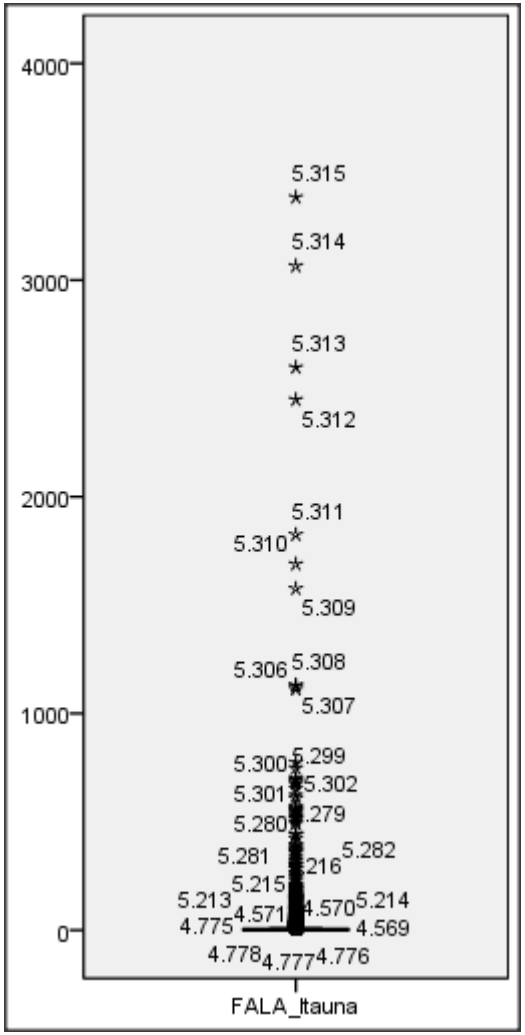
Caracterização dos corpora

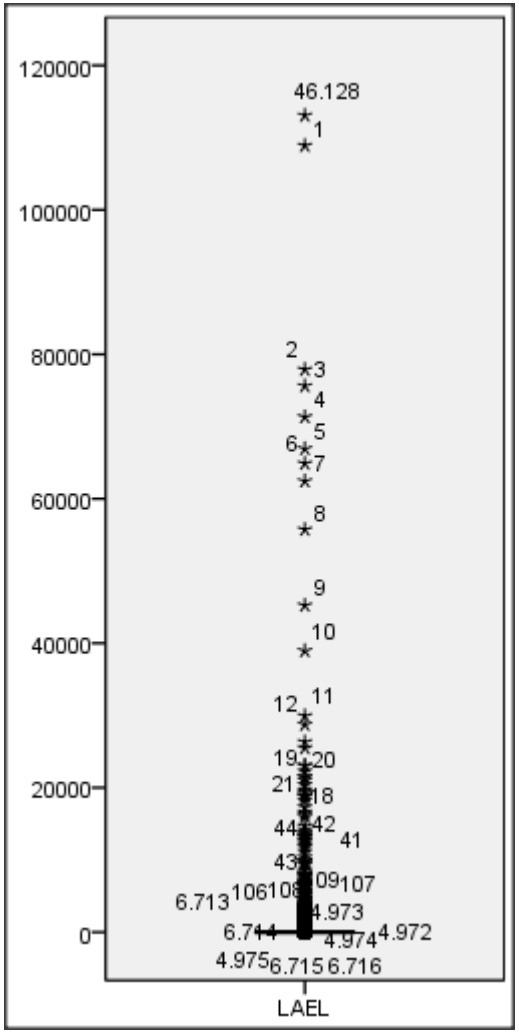
CORPUS	FALA_Itaúna	JORNAIS_Itaúna	LAEL
n	76.027	1.997.942	2.892.505
palavras diferentes	5.315	66.971	46.128
comunidade de fala	igual	igual	diferente
tipo de registro	fala	escrita	fala
tamanho do corpus	pequeno	grande	grande

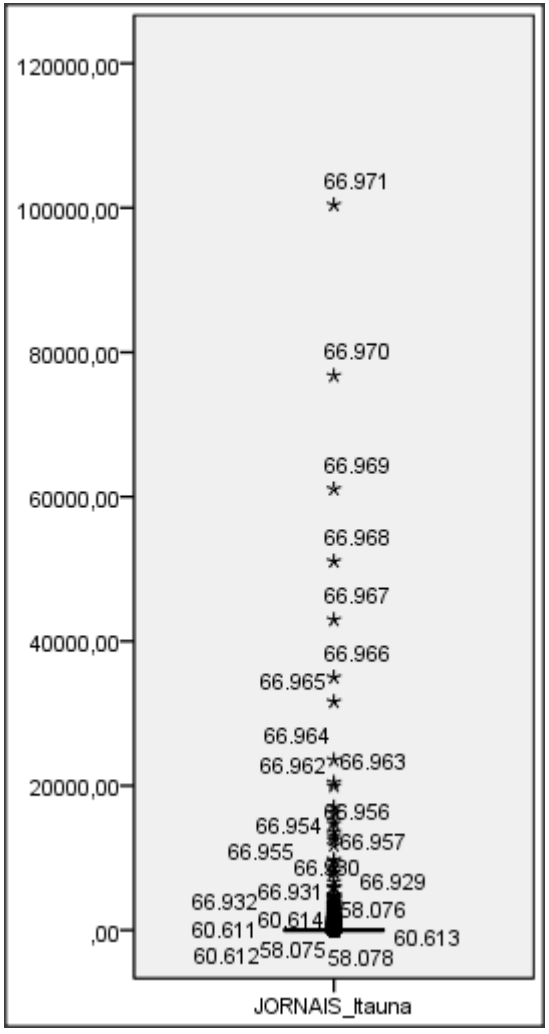
Estatística descritiva (dispersão)

Corpus	N	Mínimo	Máximo	Média	Variância
LAEL	46.128	1	113.061	62,7	1646711,8
FALA_Itauna	5.315	1	3.382	14,4	10839,6
JORNAIS_Itauna	66.971	1	100.412	29,8	459858,1

Boxplot (distribuição dos dados)







Método para análise comparativa

- Os dados são muito dispersos.

- Não convém utilizar o método convencional de correlação: *correlação de Pearson*
  - A correlação de Pearson é muito sensível a valores muito dispersos (cf. Pagano e Gauvreau, 2004)

- Proposta: **correlação de postos de Spearman**
  - Cria-se uma espécie de ranqueamento dos dados, minimizando o efeito dos valores atípicos.

Resultado da comparação entre o corpus de fala e os outros corpora

FALA	LAEL	JORNAIS
Coeficiente de correlação de Spearman	0,672	0,570
Significância	<0,001	<0,001
Percentual de correlação	67%	57%

Conclusões

- Há correlação estatisticamente significativa entre os corpora de fala de Itaúna, de fala do LAEL e de escrita de Itaúna/MG.
- Corpora de fala de diferentes dialetos do PB podem ser comparados (correlação de 67%)
- Corpora de diferentes registros (fala e escrita) de um mesmo dialeto podem ser comparados (correlação de 57%)
- A frequência das palavras observada em corpora de tamanhos mais significativos pode servir de referência para determinar a frequência das palavras em uma comunidade específica.
- Pelo método, é possível ranquear diferentes corpora: a correlação entre FALA e LAEL é maior do que entre FALA e JORNAIS.

Referências

- BYBEE, J. *Phonology and language use*. Cambridge: Cambridge, 2001.
- LABOV, Willian. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press, 1972.
- LAEL. *Banco de dados do português*. São Paulo: PUC/SP. Disponível em: <http://www2.lael.pucsp.br/corpora/>.
- OLIVEIRA, Alan Jardel. *Variação em itens lexicais terminados em /l/ V na cidade de Itaúna/MG*, 2006. 211 f. Dissertação (Mestrado em Estudos Lingüísticos) - Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2006.
- OLIVEIRA, Alan Jardel . *Velarização da lateral alveolar no falar de Itaúna/MG*. In: VIEGAS, Maria do Carmo. (Org.). Minas é plural. Belo Horizonte: UFMG, 2011.
- PAGANO, M. e GAUVREAU, K.. *Princípios de Bioestatística*. Ed. Thomson, 2ª Edição, SP, 2004.
- SARDINHA, Tony Berber. *Lingüística de Corpus: histórico e problemática*. DELTA [online]. 2000, vol.16, n.2, pp. 323-367.
- VIEGAS, M. C. ; OLIVEIRA, A. J. . *Apagamento da vogal em sílaba /l/ V átona final em Itaúna/MG e atuação lexical*. Revista da ABRALIN, v. 2, 2008.
- VIEGAS, M. C. ; OLIVEIRA, A. J. . Apagamento de /l/ v em sílaba átona final em Itaúna/Minas Gerais. In: Vanderci Aguilera. (Org.). *Para a História do Português Brasileiro*: vozes, veredas, voragens. Londrina: Eduel, 2009, v. VII, p. 393-409.