

# Padrões rítmicos, domínios prosódicos e modelagem probabilística em Corpora do português

Charlotte Galves

## Resumo

O projeto interdisciplinar Padrões rítmicos, domínios prosódicos e modelagem probabilística em corpora do português tem como objetivo a obtenção de evidências estatísticas em dados de textos escritos e em dados de fala de português que dêem suporte à conjectura de que certos domínios prosódicos e/ou certa informação fonético-fonológica são relevantes na discriminação das línguas naturais em classes rítmicas.

## 1 Introdução

### 1.1 Objetivos e justificativa

Conjectura-se que as classes rítmicas, nas quais as línguas naturais são discriminadas, são caracterizadas pelo fato de certos domínios prosódicos e/ou certa informação fonético-fonológica serem ou não relevantes (cf., por exemplo, Kleinhez, 1995; entre outros). Entre as propriedades conducentes ao ritmo acentual estão a maior complexidade silábica e maior irregularidade da distribuição de vogais e consoantes, a distinção forte entre sílaba acentuada e não acentuada e, presumivelmente, um papel mais claro das fronteiras de palavra enquanto elementos delimitativos de sequências de sons; já entre

as propriedades conducentes ao ritmo silábico, estão a maior simplicidade silábica e maior regularidade da distribuição de vogais e consoantes, a não distinção entre sílaba acentuada e não acentuada no que respeita à composição e duração da sílaba e um papel demarcativo das fronteiras de constituintes prosódicos superiores à palavra, designadamente do sintagma fonológico.

A meta deste projeto é a obtenção de evidências estatísticas em corpora escritos e de fala que dêem suporte à essa conjectura. Com essa meta, de um ponto de vista prático, nosso objetivo é fazer modelos para cadeias simbólicas obtidas codificando características fonológicas de textos escritos do português brasileiro e europeu contemporâneos e do português clássico e em corpora de fala do português brasileiro e europeu contemporâneos.

A codificação será feita marcando características fonológicas conjecturadas como sendo relevantes para a implementação do ritmo da língua (por exemplo, a distribuição de consoantes (Cs) e vogais (Vs), as fronteiras silábicas, as fronteiras de palavra prosódica, o acento de palavra), utilizando as funcionalidades da ferramenta computacional FreP, em desenvolvimento pelo Centro de Pesquisa Onset.

Os modelos pesquisados deverão satisfazer o princípio de Comprimento Mínimo de Descrição (Minimum Description Length MDL), introduzido por Jorma Rissanen nos anos 70. Isto é, trata-se de encontrar o modelo, numa classe de modelos dada, que minimize a soma dos seguintes comprimentos: comprimento da sua descrição e comprimento da compressão da cadeia simbólica comprimida usando o modelo. O segundo ponto é equivalente a pedir que o modelo maximize a verosimilhança da seqüência.

Uma classe de cadeias simbólicas que tem se revelado suficientemente flexível e econômica e tem sido aplicada com sucesso tanto a dados biológicos quanto linguísticos são as Cadeias de Memória Variável. Essas cadeias são descritas por árvores de contextos probabilísticas.

Há atualmente algoritmos, como o algoritmo Contexto ou o algoritmo das Florestas probabilísticas, que permitem estimar de maneira consistente as árvores de contextos probabilísticas subjacentes a uma cadeia simbólica dada. Isso significa o seguinte: se a cadeia for efetivamente gerada utilizando-se como mecanismo de geração uma árvore de contexto probabilística, e se a amostra for suficientemente grande, então o algoritmo de estimação identifica exatamente essa árvore. Acrescente-se a isso que há evidências empíricas de que o algoritmo das Florestas probabilísticas é também robusto.

Isso significa o seguinte: se uma cadeia for engendrada majoritariamente por uma árvore probabilística dada, e em seguida sofrer pequenas sub-sequências

simbólicas engendradas por outras árvores, ou simplesmente perturbadas por ruído na manipulação dos dados, então o algoritmo das Florestas probabilísticas identifica aparentemente a árvore majoritariamente responsável pela geração dos dados. Essa propriedade do algoritmo está em fase de demonstração.

Um dos principais interesses das árvores de contexto probabilísticas é que elas podem ser interpretadas linguisticamente. Isso abre um campo de interação frutuosa entre matemáticos e linguistas. Com efeito, os contextos probabilísticos definem em cada passo a porção do passado pertinente para a escolha do próximo símbolo. Em proteômica, é um fato experimentalmente comprovado que os contextos probabilísticos descrevem bem domínios biológicos na cadeia de amino-ácidos definindo uma proteína. Trata-se então de entender se em linguística também a noção de contexto probabilístico fornece uma formalização adequada para os domínios prosódicos ou a informação prosódica relevantes. Se assim for, teremos encontrado uma ferramenta estatística para detectar características prosódicas relevantes em cadeias simbólicas obtidas codificando-se corpora escritos.

A distribuição das tarefas entre linguistas e matemáticos fica agora clara.

1. Em primeiro lugar cabe aos linguistas propor as questões e conjecturas relevantes a serem tratadas.
2. Segundo, a partir dessas questões cabe aos linguistas propor a codificação pertinente dos dados linguísticos.
3. Cabe aos matemáticos propor classes de modelos que potencialmente têm as características desejáveis, e também propor métodos estatísticos para identificação de modelos que melhor se ajustem aos dados empíricos codificados.
4. Cabe aos linguistas identificar o possível significado linguístico das características dos modelos estimados em 3.
5. Cabe aos matemáticos identificar um quadro matemático mais geral no qual as características do tipo das identificadas em 4. sejam verificadas.

O item 5 explica porque um projeto interdisciplinar como o nosso pode ser interessante, de um ponto de vista estritamente matemático. Com efeito a construção de modelos probabilísticos, capazes de descrever bem o comportamento dos dados linguísticos e as propriedades desses modelos, podem sugerir

questões matemáticas novas, interessantes independentemente da sua motivação linguística inicial. Esse quadro matemático mais abstrato é também interessante para os linguistas na medida em que ele põe em evidência as características essenciais dos modelos linguísticos e os seus possíveis desenvolvimentos.

## 2 Fundamentação

### 2.1 Probabilidade e Lingüística

O presente projeto tem como ponto de partida a constatação de que o desempenho linguístico, embora submetido a restrições possivelmente categóricas de ordem gramatical, tem características típicas de um fenômeno estocástico. Isso se manifesta em particular na produção e na percepção de contornos rítmicos na fala e na escrita. Não há evidências de que haja regularidades determinísticas correspondendo a *padrões rítmicos* na fala. A própria noção de *acento secundário*, crucial na implementação do ritmo em qualquer variante do Português, parece não ter um correlato acústico caracterizável de forma booleana, embora ela seja suportada por experiências perceptuais reprodutíveis.

Isso sugere que o que caracteriza contornos rítmicos não são funções booleanas, e sim distribuições de probabilidades no espaço das sequências simbólicas, codificando os contornos acentuais ou melódicos. Ou seja, contornos acentuais parecem se comportar como processos estocásticos, cujas regularidades devem ser procuradas ao nível de suas leis probabilísticas (cf. Pierrehumbert 2003).

A utilização de idéias probabilísticas em Lingüística não pode ser considerada uma novidade. Com efeito, em 1905, Markov introduziu a classe de processos estocásticos que vieram a ser conhecidos como *Cadeias de Markov* especificamente para modelar as sequências de consoantes e vogais no poema *Eugênio Onegin* de Púshkin.

Kolmogorov em pessoa escreveu vários textos científicos a partir de 1960 sobre a modelagem do ritmo na poesia russa. Em um artigo inédito de 1962, Kolmogorov mostra evidências empíricas de que na poesia russa a omissão de sílabas tônicas no primeiro e terceiro pés de um octassílabo iâmbico são eventos independentes e identicamente distribuídos.

Na mesma época os trabalhos notáveis de Rabiner e colaboradores pro-

puseram diversos modelos probabilísticos, entre os quais as *Cadeias de Markov Ocultas*, para descrever a produção de uma sequência de fonemas constituindo palavras ou frases. Variantes desse modelo foram, em seguida, amplamente utilizadas em diversos algoritmos de identificação de fala e são até hoje a base da chamada *Engenharia Lingüística*.

Dentro da Lingüística teórica idéias probabilísticas têm sido utilizadas sistematicamente em sociolinguística, desde os trabalhos pioneiros de William Labov. Em Linguística histórica, os artigos notáveis de Anthony Kroch têm esclarecido de maneira original a relação entre a análise estatística de textos e a interpretação linguística. Essas idéias ressurgiram com muita força recentemente, associadas à proposta da chamada *Fonologia Probabilística* de Janet Pierrehumbert .

Esses desenvolvimentos podem ser talvez melhor entendidos na perspectiva da Mecânica Estatística. O paradigma da Mecânica Estística, formulado por Boltzmann, no final do século XIX, lançou as bases para um novo quadro conceitual no qual pode ser modelado e interpretado o comportamento de sistemas complexos. Do ponto de vista matemático esse quadro conceitual é a Teoria da Probabilidades. Esse quadro tem sido utilizado de forma crescente no estudo de diversos tipos de sistemas evolutivos em áreas como Biologia, Epidemiologia, Sociologia, Finanças, etc, além da Lingüística.

Em Lingüística, além de nossa própria contribuição, deve-se destacar o esforço pioneiro de reflexão na área desenvolvido pelo Instituto de Estudos da Complexidade de Santa Fé nos anos 90, o recente projeto de pesquisa *Dynamics and Metastability in phonological grammar*, coordenado por Janet Pierrehumbert e contemplado em 2002 com um apoio importante da Fundação James S. McDonnell, e os simpósios dos Meetings anuais da Sociedade Linguística da América de 2001 e 2003 dedicados à Teoria da Probabilidade em Linguística, culminando com a publicação do livro *Probabilistic Linguistics* pelo MIT Press em 2003.

A seguir apresentaremos detalhadamente os diversos aspectos do presente projeto.

## 2.2 A conjectura das classes rítmicas

A modelagem dos padrões rítmicos em línguas naturais é uma questão na fronteira da pesquisa em lingüística. A própria hipótese da existência de classes rítmicas separando as línguas naturais em grandes grupos ( [?], [?] e [?]), embora corroborada por evidências de caráter psico-lingüístico ( [?]),

não encontrava at recentemente suporte nos dados fontico-acústicos.

Uma primeira evidência acústica foi trazida em 1998 pelo artigo de Ramus, Nespore e Mehler (1999) [?]. Este artigo mostrou evidências que medidas empíricas do tempo relativo ocupado pelas vogais e a variância dos comprimentos dos grupos consonantais separavam um conjunto piloto de línguas em três grandes grupos. A abordagem apresentada em [?] depende de uma marcação manual prévia dos intervalos vocálicos e consonantais. Esta tarefa consome muito tempo e depende de decisões difíceis de serem feitas de forma homogênea em larga escala.

Uma nova abordagem para o problema é apresentada em [?]. Em vez de estudar durações de intervalos vocálicos e consonantais a proposta é estudar os valores de uma função que mede, em cada instante, a “sonoridade” local do sinal acústico. O cálculo da sonoridade é feito automaticamente pelo programa Piccolo desenvolvido por Galves e Garcia. Cuesta-Albertos, J.A., Fraiman, R., Galves, A., García, J. and Svarc, M. (2007) ([?]) mostram que a função de sonoridade pode ser usada para discriminar as línguas naturais em classes rítmicas. Também há evidências empíricas de que a função sonoridade pode ser bem modelada por uma cadeia de ordem infinita (cf Fernández, Ferrari e Galves 2001 para uma apresentação atualizada da área). O desenvolvimento da teoria estatística para esta classe de processos é um tema de grande atualidade. Em Collet, Duarte e Galves (2003) é introduzido um novo procedimento de reamostragem sequencial para cadeias de ordem infinita. Em Cassandro, Collet, Galves e Garcia (2003) estuda-se a questão da estimação da fronteira de uma cadeia quantizada. Esta questão teórica nos foi sugerido pela análise do comportamento empírico da sonoridade da fala.

### **2.3 Correlatos de ritmo em textos escritos de Português Brasileiro e Europeu Moderno**

Neste projeto estamos propondo uma abordagem para detectar ritmo em textos escritos utilizando Cadeias de Markov de alcance variável. Estudamos textos do século XX de autores brasileiros e portugueses e textos históricos de escritores nascidos em Portugal entre o século XVI e XIX do Corpus histórico Tycho Brahe (cf. seção 3.1 deste projeto). Nestes textos marcamos todas as sílabas que são tônicas ou átonas, início de palavra fonológica e ponto final. Usando Cadeias de Markov de Alcance Variável estimamos padrões

para cada texto. Este método nos permite discriminar completamente entre Português Europeu Moderno e Português Brasileiro. Para os textos clássicos, observamos diferenças de ambas as línguas modernas bem como uma maior variação nos padrões atestados.

## 2.4 O modelo probabilístico

O modelo considera cada texto como uma amostra finita de uma Cadeia de Markov de Alcance Variável (VLMC, do inglês “Variable Length Markov Chain”)  $(X_i)_{i \in \mathbf{Z}}$  com valores em um alfabeto finito  $\mathcal{A} = \{0, 1, 2, 3, 4\}$  e ordem máxima que pode ser finita ou infinita. Uma VLMC é simplesmente uma Cadeia de Markov apresentada de maneira parcimoniosa como se segue.

Seja  $p : \mathcal{A} \times \mathcal{A}^K \rightarrow [0, 1]$  a probabilidade de transição da cadeia

$$p(x_0 | x_{-K}^{-1}) := \mathbf{P}(X_0 = x_0 | X_{-K}^{-1} = x_{-K}^{-1}). \quad (1)$$

A função  $c : \mathcal{A}^\infty \rightarrow \cup_{m=0}^\infty \mathcal{A}^m \cup \mathcal{A}^\infty$  é dita ser a *função contexto* do processo se

$$c : x \mapsto x_{-\ell(x)+1}^{-1} \quad (2)$$

para  $x = x_\infty^{-1}$  onde  $\ell : \mathcal{A}^\infty \rightarrow \mathbf{N} \cup \{0\}$  é o comprimento do contexto e é dado por

$$\ell(x) = \min\{k; p(a|x) = p(a|x_{-k+1}^{-1}), \text{ for all } a \in \mathcal{A}\}. \quad (3)$$

Cada valor da função contexto  $c$  pode ser representada por um ramo de uma árvore construída da seguinte forma:

- A raiz se localiza no topo;
- Ramos crescem para baixo;
- Cada nó interno têm, no máximo,  $|\mathcal{A}|$  descendentes.;
- O contexto  $w = c(x_\infty^{-1})$  é representado por um ramo, cujo sub-ramos superior é determinado por  $x_{-1}$ , o sub-ramo seguinte é determinado por  $x_{-2}$  e assim por diante;
- Se  $\ell(x_\infty^{-1}) < \infty$ , o sub-ramo terminal é determinado por  $x_{-\ell(x_\infty^{-1})+1}$ .

Bühlmann and Wyner (1999) propõe um algoritmo para estimar a função contexto e conseqüentemente a árvore contexto o qual é consistente se a ordem da cadeia é limitado. Galves, Garcia e Peixoto (preprint, 2003) mostram que este algoritmo também é consistente para o caso onde a ordem da cadeia é infinita.

Até o presente momento, utilizamos o algoritmo proposto por Bühlmann e Wyner para estimar as funções contexto para cada texto separadamente usando o software R ([www.r-proj.org](http://www.r-proj.org)).

## 2.5 Índícios rítmicos em Português Europeu e Brasileiro

Até o presente momento lidamos com cada texto separadamente. Entretanto, acreditamos que existam certas características no ritmo que caracterizam e discriminam entre Português Europeu e Brasileiro. Neste caso, postularemos árvores típicas para cada língua baseadas nas árvores estimadas e daí aplicaremos teste da razão de verossimilhança considerando o caso de hipótese nula simples

$$H_0 : \quad \tau = \tau_0 \quad (c = c_0) \quad (4)$$

$$H_a : \quad \tau = \tau_1 \quad (c = c_1) \quad (5)$$

onde  $c_0$  e  $c_1$  são árvores contexto fixas e  $\tau_0$  e  $\tau_1$  são as árvores contexto correspondentes. Assuma também que  $c_1 \geq c_0$ . Considere a estatística do teste:

$$X_n^2 = -2 \sum_{k=1}^m \log \left( \frac{\hat{p}_{c_0}(X_n^k)}{\hat{p}_{c_1}(X_n^k)} \right) \quad (6)$$

onde  $\hat{p}_{c_i}(X_n^k)$  é a estimativa de máxima verossimilhança do modelo com função contexto  $c_i$ . Defina

$$r = (|\tau_{c_1}| - |\tau_{c_0}|)(|\mathcal{A}| - 1) \quad (7)$$

Sob a hipótese nula

$$X_n^2 \rightarrow \chi_r^2 \quad (8)$$

em distribuição quando  $n \rightarrow \infty$ .

Como um caso particular do teorema acima temos o teste de hipótese para o caso:

$$H_0 : \quad \tau = \tau_j \quad (c = c_j)$$

$$H_a : \quad \tau \neq \tau_j \quad (c \neq c_j)$$

onde  $c_{BP}$  e  $c_{EP}$  são as árvores contexto postuladas para Português Brasileiro e Português Europeu Moderno respectivamente. Neste caso,  $c_1 = c^*$  é a função contexto completa. Note que neste caso, completa não significa a árvore com todos os contextos, visto que alguns padrões são impossíveis de serem observados na língua portuguesa.

## 2.6 Testando diferenças entre duas populações

Um problema completamente em aberto que pretendemos abordar é como testar se duas línguas tem diferentes funções contexto sem termos que postular uma função contexto a priori. Neste caso, podemos ter que duas populações tenham a mesma função contexto mas diferentes probabilidades de transição. Este caso não é coberto pela teoria tradicional de teste de hipótese, pois não há, a princípio, uma estatística do teste que discrimine as hipóteses que esteja baseada em uma métrica.

# 3 Material e Metodologia

## 3.1 Material

O material a ser utilizado no desenvolvimento desta pesquisa refere-se a um conjunto de textos escritos de português brasileiro e europeu contemporâneos e de português clássico.

O conjunto de textos de português brasileiro e europeu contemporâneos consiste, respectivamente, em textos jornalísticos constantes da coleção das edições completas dos anos de 1994 e 1995 dos jornais Folha de São Paulo (Brasil) e Público (Portugal). Esta coleção foi compilada pela Linguateca (centro de recursos para o processamento computacional da língua portuguesa e que permite o acesso, via internet, a recursos já existentes, como, por exemplo, corpora lingüísticos de língua portuguesa já digitalizados, cf.

<http://www.linguateca.pt>) e ocorreu no quadro do CLEF (Cross-Language Evaluation Forum, Forum de avaliação entre várias línguas). O corpus completo que consta dos referidos textos jornalísticos pode ser obtido em formato de texto (extensão .txt) pela solicitação do mesmo através do endereço eletrônico: [http://acdc.linguateca.pt/aval\\_conjunta/CLEF/CHAVE/](http://acdc.linguateca.pt/aval_conjunta/CLEF/CHAVE/).

Quanto ao conjunto de textos de português clássico, tais textos se referem ao conjunto de textos constantes do Corpus Histórico do Português

Tycho Brahe (<http://www.ime.usp.br/tycho>). O Corpus Histórico do Português Tycho Brahe é um corpus eletrônico anotado, composto de textos portugueses escritos por autores nascidos entre 1435 e 1835. Atualmente, 48 textos (2.279.455 palavras) estão disponíveis para pesquisa livre, com um sistema de anotação lingüística em duas etapas: anotação morfológica (aplicada em 23 textos); e anotação sintática (aplicada em um texto). O Corpus Tycho Brahe é desenvolvido junto ao projeto temático Padrões Rítmicos, Fixação de Parâmetros e Mudança Gramatical II (projeto temático financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo, processo Fapesp 04/03643-0). O conjunto de textos constantes do Corpus Tycho Brahe pode ser obtido através de solicitação, pelo acesso ao endereço eletrônico: <http://www.ime.usp.br/tycho/corpus/>.

Cabe acrescentar que, eventualmente, serão acrescentados dados de fala do português brasileiro e europeu contemporâneos, a depender do desenvolvimento da pesquisa.

## 3.2 Metodologia

A metodologia a ser empregada nesta pesquisa consiste, primeiramente, na codificação lingüística dos dados de português brasileiro e europeu contemporâneos e de português clássico, através da marcação de características fonológicas conjecturadas como sendo relevantes para a implementação do ritmo da língua (por exemplo, a distribuição de consoantes (Cs) e vogais (Vs), as fronteiras silábicas, as fronteiras de palavra prosódica, o acento de palavra), utilizando as funcionalidades da ferramenta computacional FreP, em desenvolvimento pelo Centro de Pesquisa Onset.

FreP é uma ferramenta eletrônica que permite a extração de informação de frequência de unidades fonológicas em português europeu, no nível da palavra prosódica e abaixo destes níveis. A ferramenta FreP vem sendo desenvolvida por Marina Vigário, Fernando Martins e Sónia Frota da Universidade de Lisboa (Portugal) no âmbito do projeto Padrões de Frequência na Fonologia do Português - Investigação e Aplicações, processo PTDC/LIN/70367/2006.

O ajuste dos modelos probabilísticos (via VLMC cf. seção 2.4) dos dados obtidos a partir da aplicação do FreP será realizado através do algoritmo PST e outros que serão desenvolvidos ao longo desta pesquisa.

## 4 Equipe

A equipe de pesquisadores principais do presente projeto é apresentada abaixo:

1. Charlotte Marie Chambelland Galves (coordenadora geral do projeto) - Lingüística (Unicamp)
2. Jefferson Antonio Galves - Estatística (USP)
3. Maria Bernadete Marques Abaurre - Lingüística (Unicamp)
4. Luciani Ester Tenani - Lingüística (Unesp, São José do Rio Preto)
5. Flaviane Romani Fernandes Svartman - Lingüística (Unicamp)
6. Nancy Lopes Garcia - Estatística (Unicamp)
7. Jesús Enrique García - Estatística (Unicamp)
8. Verónica A. González-López - Estatística (Unicamp)
9. Arnaldo Mandel - Computação (USP)
10. Denise Duarte - Estatística (UFMG)
11. Sónia Marise de Campos Frota (colaboradora externa) - Lingüística (Universidade de Lisboa, Portugal)
12. Marina Cláudia Vigário (colaboradora externa) - Lingüística (Universidade de Lisboa, Portugal)
13. Marzio Cassandro (colaborador externo) - Física (La Sapienza, Roma)
14. Pierre Collet (colaborador externo) - Matemática (CNRS, Paris)
15. John Goldsmith (colaborador externo) - Lingüística (Universidade de Chicago)

## 5 Cronograma de atividades

Este projeto procurará seguir o cronograma especificado abaixo:

## 5.1 Primeiro ano de projeto

1. Codificação das características fonológicas relevantes (a distribuição de Cs e Vs, as fronteiras silábicas, as fronteiras de palavra prosódica, o acento de palavra) nos textos constantes dos corpora do jornal O Público de 1994 e 1995 e do Corpus Tycho Brahe, com recurso ferramenta FreP.
2. Levantamento das questões relevantes sobre as propriedades rítmicas do português europeu atual e do português clássico, bem como o estabelecimento de uma hipótese da deriva histórica do ritmo do português.
3. Modelagem estocástica das sequências simbólicas linguisticamente codificadas.
4. Interpretação linguística dos resultados da modelagem estocástica, designadamente tendo em consideração as questões e hipóteses referidas em 2.
5. Elaboração de artigos, constando dos resultados obtidos, a serem apresentados em encontros científicos e a serem submetidos a revistas científicas das especialidades envolvidas.
6. Realização do primeiro workshop do presente projeto, no qual estará reunida toda a equipe de pesquisadores e no qual serão apresentados e discutidos (i) os principais resultados obtidos no decorrer do primeiro ano de pesquisa e (ii) os rumos desta para o segundo ano.

## 5.2 Segundo ano de projeto

1. Codificação das características fonológicas relevantes (a distribuição de Cs e Vs, as fronteiras silábicas, as fronteiras de palavra prosódica, o acento de palavra) nos textos constantes do corpus do jornal Folha de São Paulo de 1994 e 1995, com recurso versão da ferramenta FreP, adaptada para português brasileiro.
2. Acréscimo de corpora de fala de português brasileiro e europeu contemporâneos e codificação das características fonológicas relevantes (a distribuição de Cs e Vs, as fronteiras silábicas, as fronteiras de palavra prosódica, o acento de palavra) nestes corpora.

3. Levantamento e desenvolvimento aprofundado das questões relevantes sobre as propriedades rítmicas do português europeu e brasileiro atuais e do português clássico, bem como a investigação da hipótese elaborada no primeiro ano sobre a deriva histórica do ritmo do português.
4. Modelagem estocástica das sequências simbólicas linguisticamente codificadas no corpus do jornal Folha de São Paulo e nos eventuais corpora de fala acrescentados.
5. Interpretação linguística dos resultados da modelagem estocástica, designadamente tendo em consideração as questões e hipóteses referidas em 3.
6. Elaboração de artigos, constando dos resultados obtidos, a serem apresentados em encontros científicos e a serem submetidos a revistas científicas das especialidades envolvidas.
7. Realização do segundo workshop do presente projeto, no qual estará reunida toda a equipe de pesquisadores e no qual serão apresentados e discutidos os principais resultados obtidos ao longo de toda a pesquisa.

## References

- [1] Abercrombie, D., (1967). *Elements of general phonetics*, Chicago: Aldine.
- [2] Cassandro, M., Collet, P., Duarte, D., Galves, A., and Garcia, J. (2005). Tied quantized chains and cross-linguistic estimation of the cut-points of the speech sonority. *Manuscript*.
- [3] Cros A., Demolin D, and Flesia G, (2005). On the relationship between intra-oral pressure and speech sonority, Paper presented at *Interspeech 2005- Eurospeech*, Lisbon. (can be downloaded from [www.ime.usp.br/~galves/artigos/sonopres.eps](http://www.ime.usp.br/~galves/artigos/sonopres.eps)).
- [4] Cuesta-Albertos, J. A., Fraiman, R. and Ransford, T. (2004). A sharp-form of the Cramér-Wold theorem. *Manuscript*
- [5] Duarte, D, Galves, A., Lopes, N. and Maronna, R.(2001). The statistical analysis of acoustic correlates of speech rhythm. Paper presented at the *Workshop on Rhythmic patterns, parameter setting and language change*, ZiF, University of Bielefeld. Can be downloaded from <http://www.physik.uni-bielefeld.de/complexity/duarte.eps>
- [6] Galves, A., Garcia, J., Duarte, D. and Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. Paper presented at *Speech Prosody 2002*, Aix-en-Provence (can be downloaded from [www.lpl.univ-aix.fr/sp2002/pdf/galves-etal.eps](http://www.lpl.univ-aix.fr/sp2002/pdf/galves-etal.eps)).
- [7] Lloyd, J. (1940). *Speech signal in telephony*, London.
- [8] Mehler, J.; Dupoux, E.; Nazzi, T.; Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: the infant's viewpoint. *Signal to syntax: bootstrapping from speech to grammar in early acquisition*, J.L. Morgan and K. Demuth, eds.
- [9] Piccolo can be downloaded from [www.ime.usp.br/~tycho/prosody/piccolo](http://www.ime.usp.br/~tycho/prosody/piccolo)
- [10] Pike, K.L. (1945). *The intonation of American English*, Ann Arbor: University of Michigan Press.

- [11] Praat program and manuals. Can be downloaded from [www.praat.org](http://www.praat.org).
- [12] Ramus, F. (2002) Acoustic correlates of linguistic rhythm: perspectives. *Speech Prosody 2002*, Aix-en-Provence. Can be download from [www.lpl.univ-aix.fr/sp2002/pdf/ramus.eps](http://www.lpl.univ-aix.fr/sp2002/pdf/ramus.eps).
- [13] Ramus, F., Nespors, M. and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, **73**, 265-292.