Comportamento estocástico, fenômenos críticos e identificação de padrões rítmicos nas línguas naturais

Projeto de Núcleo de Excelência

29 de Setembro de 2003

Conteúdo

1 Sumário do Projeto

Comportamento estocástico, fenômenos críticos e identificação de padrões rítmicos nas línguas naturais

Este projeto interdisciplinar tem como objetivo principal desenvolver a Teoria dos Processos Estocásticos, para formular rigorosamente e tratar os seguintes problemas centrais da Lingüística:

1. a questão da existência ou não de padrões rítmicos nas línguas naturais;
2. a existência de uma tipologia discreta caracterizada por pontos críticos bem definidos, em oposicão a um contínuo rítmico;
3. a existência de marcas do ritmo no sinal acústico de fala e em textos escritos.

Além de desenvolver resultados matemáticos originais interessantes por si só, o projeto usará o quadro conceitual da Teoria das Probabilidades para efetivamente interpretar, usando a análise estatística, os dados linguísticos, visando obter uma compreensão mais profunda das questões formuladas. Como sub-produto o projeto desenvolverá ferramental estatístico e computacional necessário ao tratamento dos dados linguísticos. Esse último aspecto aponta para a possibilidade de desdobramentos tecnológicos na área de Engenharia da Linguagem.

Para atingir seus objetivos a equipe do Projeto realizará os seguintes tipos de atividades:

1. pesquisa matemática para estudar as propriedades dos diversos modelos formais propostos pelo projeto;
2. pesquisa lingüística para atualizar e reformular as questões centrais do projeto à luz dos novos resultados matemáticos;
3. construção de corpora de fala, acústicos e escritos, e tratamento laboratorial das amostras lingüísticas;
4. análise estatística dos dados linguísticos, para ajustar os modelos matemáticos e para encontrar evidências apoiando ou contrariando as previsões dos modelos;

5. desenvolvimento de novas ferramentas computacionais e estatísticas para análise dos dados lingüísticos.

Essas atividades de pesquisa serão feitas no Núcleo de Modelagem Estocástica e Complexidade da USP (Numec), no IME-USP, IMEC-UNICAMP, IEL-UNICAMP, Matemática-UFG, além das instituições de pesquisa estrangeiras (Laboratoire de Mathématiques- Rouen, Centre de Physique Théorique-CNRS e Ecole Polytechnique de Palaiseau, Centre de Physique Théorique-CNRS Luminy, Fisica-Roma *"La Sapienza"*, Linguistica-Ferrara, Lingüística-Lisboa, Lingüística-Braga, Cognitive Sciences-UPenn, Laboratoire de Sciences Cognitives-Ecole de Hautes Etudes en Sciences Sociales e ENS, Lingüística Bielefeld e Freiburg, Linguistics-Northwestern University). Isso torna indispensável visitas de trabalho regulares dos membros da equipe às diversas instituições hospedando o projeto.

A coordenação do trabalho será feita no Numec-USP, onde serão instalados o Laboratório de Fonética Experimental e o Labatório de Cálculo Científico do projeto.

Este projeto requer conhecimentos e competências científicas variadas envolvendo Matemática, Computação, Estatística e Lingüística. Essa formação interdisciplinar não é atualmente fornecida por nenhuma universidade brasileira. Nos últimos anos nossa equipe tem-se empenhado em enfrentar essa lacuna, propondo estágios de pesquisa e projetos de Iniciação Cientıfica, Mestrado, Doutoramento e Pós-Doutoramento em temas interdisciplinares associados à nossa pesquisa, muitos deles com apoio da FAPESP e do CNPq. Continuaremos esse trabalho de formação no quadro do presente projeto.

2 Summary of the Project

Stochastic behavior, critical phenomena and rhythmic patterns identification in natural languages

The aim of this interdisciplinary project is to develop the area of Stochastics Processes Theory in order to rigorously address the following central problems in Linguistics:

1. the question of the existence of rhythmic patterns in natural languages
2. the existence of a discrete typology characterized by well-defined critical points, as opposed to a rhythmic continuum
3. the existence of rhythmic features in the acoustic signal and in written texts.

Besides obtaining new mathematical results which are interesting by themselves, the project will use the conceptual framework of Probability Theory to effectively interpret linguistic data. The use the of statistical analysis will be crucial in order to arrive at a deeper understanding of the linguistic issues. As a by-product, the project will develop the statistical and computational tools necessary to treat the relevant linguistic data. This last aspect opens the possibility of technological developments in Language Engineering.

In order to achieve its goals, the project team will engage in the following activities:

1. mathematical research to study the properties of the formal models proposed by the project;
2. linguistic research to update and reformulate the central questions of the project in the light of the new mathematical findings;
3. acoustical and written speech corpora building, and laboratorial treatment of linguistic samples;
4. statistical analysis of linguistic data in order to adjust the mathematical models and to find supporting or negative evidence for the predictions of the models;
5. development of new computational and statistical tools for the analysis of linguistic data.

The reseach activities will be conducted at the Núcleo de Modelagem Estocástica e Complexidade da USP (Numec), at IME-USP, IMEC-UNICAMP, IEL-UNICAMP, Mathematics-UFMG, and also at foreign research institutions (Laboratoire de Mathématiques- Rouen, Centre de Physique Théorique-CNRS e Ecole Polytechnique de Palaiseau, Centre de Physique Théorique-CNRS Luminy, Fisica-Roma *"La Sapienza"*, Lingüística-Ferrara,Lingüística-Lisboa, Lingüística-Braga, Cognitive Sciences-UPenn, Laboratoire de Sciences Cognitives-Ecole de Hautes Etudes en Sciences Sociales e ENS, Lingüístics Bielefeld and Freiburg, Linguistics-Northwestern University). Therefore regular working visits of members of the project team at the different institutions that host the project will be absolutely necessary.

Work coordination will be at Numec-USP, where the Experimental Phonetics Laboratory and the Scientific Calculus Laboratory will be located.

This Project requires an array of specific knowleges and competences which comprise Mathematics, Computer Science, Statistics and Linguistics. Such interdisciplinary background is not provided by any Brazilian university, at the present moment. In the last years, the project team has made a great effort to fulfill this gap by offering research training and by advising undergraduate scientific projects, M.A., Ph.D and Post-doctoral projects on interdisciplinary topics related to our research. Many such projects have been funded by FAPESP and CNPq. We shall continue this researchers' training program within the framework of the present project.

3 Projeto de Pesquisa

3.1 Introdução

Este projeto interdisciplinar tem como objetivo principal o desenvolvimento de uma nova área de pesquisa matemática que poderia ser adequadamente chamada de *Fonologia Probabilística do Ritmo*. Seu escopo é o estudo das propriedades matemáticas das cadeias de ordem infinita descrevendo os contornos acentuais em línguas naturais, e mais geralmente dos processos estocásticos descrevendo a interface entre a sintaxe e o ritmo no desempenho lingüístico. Utilizaremos o quadro conceitual da Teoria das Probabilidades e o paradigma da Mecânica Estatística para definir e identificar características rítmicas das línguas naturais.

Os resultados produzidos devem responder a perguntas fundamentais da Lingüística, a saber:

1. a questão da existência ou não de padrões rítmicos nas línguas naturais;

2. a existência de uma tipologia discreta caracterizada por pontos críticos bem definidos, em oposição a um contínuo rítmico;

3. a existência de marcas do ritmo no sinal acústico de fala e em textos escritos.

Este projeto continuará o trabalho de pesquisa matemática em Teoria das Probabilidades iniciado pelo Projeto Pronex *Fenômenos Críticos em Probabilidades e Processos Estocásticos*, o Projeto Temático FAPESP *Padrões Rítmicos, Fixação de Parâmetros e Mudança Lingüística* e o Projeto CNPq (Edital 2000) *Técnicas Probabilísticas de Identificação de Padrões Rítmicos com Aplicação à Lingüística*. Esses projetos anteriores já vinham utilizando de forma continuada a Teoria das Probabilidades e o paradigma da Mecânica Estatística para modelar a interface sintaxe-fonologia. Esse esforço de pesquisa deu origem a um conjunto de resultados novos na área de Teoria das Probabilidades, com interesse matemático intrínseco. Além disso, esses projetos construiram o quadro conceitual matemático dentro do qual é possível dar um tratamento rigoroso das questões lingüísticas que deram origem à pesquisa. Esse quadro conceitual levou a predições já verificadas experimentalmente e sugeriu novas linhas de investigação. Finalmente, esses projetos produziram como sub-produto um conjunto de ferramentas estatísticas e computacionais, entre elas os programas de código aberto

Sotaq, *Vocale* e *Piccolo*. Todos esses resultados são amplamente descritos nos relatórios finais dos projetos mencionados acima (cf. Anexos 1, 2 e 3).

Em resumo, o presente projeto tem como objetivo desenvolver a Teoria dos Processos Estocásticos, para formular e tratar rigorosamente um conjunto de problemas centrais da Lingüística. Além de desenvolver resultados matemáticos originais interessantes por si só, o projeto usará o quadro conceitual da Teoria das Probabilidades para efetivamente interpretar, usando a análise estatística, os dados linguísticos, visando obter uma compreensão mais profunda das questões formuladas. Esse é exatamente o paradigma historicamente seguido pela Física. Nossa pesquisa é em particular inspirada pelas relação fértil que a Mecânica Estatística estabeleceu entre a Termodinâmica e a Teoria das Probabilidades. Como sub-produto o projeto desenvolverá ferramental estatístico e computacional necessário ao tratamento dos dados linguísticos. Esse último aspecto aponta para a possibilidade de desdobramentos tecnológicos na área de Engenharia da Linguagem.

3.2 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 determínisticas correspondendo a *padrõoes rítmicos* na fala. A própria noção de *acento secundário* (150), crucial na implementação do ritmo em qualquer variante do Português, parece nao 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 provuradas 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 Oneguin* de Püshkin.

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, recentemente publicado (99), 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 colaboradores propuseram diversos modelos probabilísticos, entre os quais as *Cadeias de Markov Ocultas*, para descrever a produção de uma sequência de fonemas constituindo palavras ou frases (cf. (**?**), (**?**)). Variantes desse modelo foram, em seguida, amplamente utilizadas em diversos algoritmos de identificação de fala e são até hoje a base da chamada *Engenharia Lingüıstica*.

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 (cf. (**?**) e também (**?**) para uma apresentação dessas idéias em outras áreas da Lingüística).

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 (78), 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 equipe do presente projeto começou em 1993 um trabalho de pesquisa sistemática diretamente inspirado pelo paradigma da Mecânica Estatística. A idéia era utilizar "estados de Gibbs" como modelos para a interface sintaxe-fonologia. Essa idía foi desenvolvida nos artigos de Collet, Galves e Lopes (1995) , Cassandro, Collet, Galves e Galves (1999) e Fernández e Galves (2000)) cujas cópias seguem em anexo e aos quais remetemos os leitores, para uma apresentação detalhada e formal do nosso modelo de base.

Esse trabalho deu origem ao Projeto Temático, FAPESP *Padrões rítmicos, fixação de parâmetros e mudança língüística*, cujos resultados estão apresentados na seção consagrada aos resultados de auxílios anteriores.

A seguir apresentaremos detalhadamente os diversos aspectos do presente projeto.

3.3 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 ((112), (136) e (2)), embora corroborada por evidências de caráter psico-lingüístico ((117)), não encontrava até recentemente suporte nos dados fonético-acústicos.

Uma primeira evidência acústica foi apresentada pelo artigo de Ramus, Nespor e Mehler (1999) , sendo o segundo autor membro colaborador deste projeto. 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 (147) 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 (63). 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, que vem sendo desenvolvido por Galves e Garcia e que pode ser obtido

livremente para pesquisas acadêmcias no endereço `http://www.ime.usp.br/~tycho/prosody`. A seguir apresentamos as questões matemáticas sugeridas por essa abordagem.

3.4 Modelagem estocástica da sonoridade da fala

Há evidências empíricas de que a função sonoridade pode ser bem modelada por uma cadeia quantizada de ordem infinita. Neste modelo a sonoridade de cada língua é controlada por uma cadeia de ordem infinita assumindo valores num alfabeto binário. Essa cadeia, cuja lei depende da língua, determina os intervalos passados nas regiões de alta e baixa sonoridade. Em seguida, condicionalmente à região na qual se encontra, a distribuição dos valores assumidos pela sonoridade segue probabilidades independentes da língua. Em consequência, todas as informações sobre o rítmo de cada língua devem estar contidas unicamente na cadeia de ordem infinita subjacente à sonoridade.

Seguem daíduas questões probabilísticas interessantes. A primeira é como estimar o ponto de corte separando as regiões de alta e baixa sonoridade. Essa questão é tratada em Cassandro, Collet, Duarte, Galves e Garcia (2003) que demonstram um teorema de consistência para uma família de estimadores do ponto de corte. O próximo passo da pesquisa ser'a estudar as flutuações desses estimadores e estimar suas variâncias. Como esses estimadores são definidos como argumentos de pontos de máximo de certos funcionais estocásticos, essa é uma questão matemática extremamente delicada.

É interessante observar que as predições feitas pelo modelo apresentado em (**?**) são comprovadas experimentalmente de forma bastante satisfatória. As consequências dessas predições para a compreensão da fonologia do ritmo é um dos temas a serem pesquisados pelo projeto.

A segunda questão derivada desse modelo é a estimação de probabilidades de cilindros de tamanho fixo, de probabilidades de transição e da entropia das cadeias de ordem infinita subjacentes. O desenvolvimento de teoria inferencial para cadeias de ordem infinita é um tema de grande importância e atualidade. Em Collet, Duarte e Galves (2003) é introduzido um novo procedimento de reamostragem sequencial para cadeias de ordem infinita e demonstrado um teorema-limite central da reamostragem justificando esse procedimento. Esse procedimento se aplica a probabilidades de cilindros de tamanho fixo. Isso abre a possibilidade de testar a igualdade das proporções de tempo

passado em regiões de alta ou baixa sonoridade em cada classe de línguas.

Em Abadi e Galves (2003) e Gabrielli, Galves e Guiol (2003) são apresentados resultados preliminares à questão da estimação da entropia em cadeias de ordem infinita. O primeiro trabalho discute a questão da velocidade de convergência da entropia de cadeias de Markov de ordem finita convergindo de forma canônica para uma cadeia de ordem infinita. No segundo artigo apresenta-se um teorema-limite central para as entropias relativas empíricas dessas cadeias aproximantes. Em nosso projeto daremos sequência a esse trabalho, pesquisando as condicções nas quais é possível estender o procedimento sequencial e o teorema-limite central de reamostragem para cadeias de ordem infinita para funcionais que dependem de toda a trajetória da cadeia como é o caso da entropia e das probabilidades de transição das cadeias.

Em todas essas pesquisas uma etapa prévia importante foi a obtenção de majorantes finos para a distância, em sentido $\bar{d}$ por exemplo, entre as leis de uma cadeia de ordem infinita e suas aproximações markovianas. Esse é um tema recorrente em nosso projeto (cf. (16), (18), (**?**)) que continuará sendo pesquisado e cuja importância matemática transcende as aplicações acima mencionadas. Outro tema recorrente é a obtenção de resultados finos para aproximações de ordem infinita (cf. Abadi e Galves 2002 para uma apresentaçào atualizada da área). Uma apresentação atualizada das cadeias de ordem infinita pode ser encontrada em Fernández, Ferrari e Galves (2001) .

3.5 Correlatos de ritmo em textos escritos de Português Brasileiro e Europeu Moderno

Além das cadeias de ordem infinita subjacentes à sonoridade, nós estudamos nesse projeto também as chamadas *Cadeias de Markov de Alcance Variável* (VLMC, do inglês "Variable Length Markov Chain"). Elas aparecem como modelos para sequências codificadas de sílabas em textos escritos.

O modelo considera cada texto como uma amostra finita de uma Cadeia de Markov de Alcance Variável com valores em um alfabeto finito e ordem máxima que pode ser finita ou infinita. Uma VLMC é simplesmente uma Cadeia de Markov apresentada de maneira parcimoniosa.

Bühlmann and Wyner (1999) propõe um algoritmo para estimar a função contexto que é consistente se a ordem da cadeia é limitado. Galves, Garcia e Peixoto (manuscrito, 2003) mostram que

este algoritmo também é consistente para o caso onde a ordem da cadeia é infinita.

Considerações lingüísticas sugerem que padrões rítmicos distintos devem corresponder a árvores de contexto distintas. Essa é a base para a nova abordagem que estamos propondo para a identificação de padrões rítmicos em textos escritos. 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. 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 as árvores de contexto e as probabilidades de transição modelando cada texto. Este método nos permite discriminar completamente entre Portugês Europeu Moderno e Português Brasileiro. Para os textos clássicos, observamos diferenças em relação a ambas as línguas modernas, bem como uma maior variação nos padrões atestados. Galves, Galves, Garcia e Peixoto(manuscrito 2003) apresenta os resultados dessa pesquisa.

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 (`http://www.r-proj.org`).

Apesar de não haver consenso entre as árvores estimadas em cada língua, 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 o teste da razão de verossimilhança para testar esta hipótese.

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, é concebível 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.

Esta é uma linha extremamente promissora de pesquisa, totalmente original do ponto de vista do procedimento de análise dos dados lingüísticos e de grande atualidade e interesse do ponto de vista da Teoria das Probabilidades. Nosso projeto desenvolverá essa direção de pesquisa intensivamente.

3.6 Uma análise otimalista do ritmo

A análise otimalista de contornos acentuais introduzida em Sandalo, Mandel, Abaurre e Galves (2002) ataca a questão da identificação de padrões rítmicos de um outro ponto de vista. Computacionalmente, esta abordagem motivou a criação do programa Sotaq, desenvolvido por Mandel, em linguagem PERL, a partir de uma primeira versão desenvolvida por Collet, Galves e Galves em linguagem C `http:www.ime.usp.br/~tycho/prosody/sotaq`. A fundamentação empírica deste ponto de vista exige o desenvolvimento de ferramentas estatísticas novas que permitam identificar restrições e hierarquias que melhor se ajustem aos dados. Estas são direções de pesquisa extremamente promissoras, tanto do ponto de vista matemático, quanto do seu potencial de aplicação e não apenas à análise de dados linguísticos.

A Teoria da Otimalidade foi introduzida por Prince e Smolensky (1993). Trata-se de um modelo representacional desenvolvido como uma alternativa aos modelos derivacionais que caratcterizaram a Fonologia Gerativa desde Chomsky e Halle (1968). A hipótese fundamental dessa nova abordagem era a existência de um conjunto universal de restri cões sobre realizações possíveis de formas lingüísticas. Nesse modelo as línguas se diferenciam pela maneira como hierarquizam essas restrícões. Formalmente , o modelo supõe a existência de duas funções, uma gerando candidatos a formas ótimas, e outra avaliando esses candidatos a partir do número de infrações que cometem às restrições. O ordenamento dos candidatos é feito utilizando um tipo de ordem lexicográfica.

A Teoria da Otimalidade foi claramente inspirada pelo paradigma da Mecânica Estatística. O número de infrações às restrições cometidas corresponde a uma função *custo* ou *energia*. O custo associado a cada infração depende da posição ocupada pela restrição na hierarquia caracterizando a língua. Se associarmos uma medida de Gibbs a essa função energia, os candidatos selecionados são aqueles que pertencem ao suporte da medida de Gibbs no limite em que a *temperatura* converge a zero (o que em Mecânica Estatística se chamaria "estados fundamentais"do sistema)

Uma abordagem otimalista bem sucedida tem como ponto de partida a consolidação do conjunto de restrições numa função única a ser minimizada - existem, de fato, várias formas de proceder a essa consolidação, e espera- se que um modelo correto para alguns problemas reflita corretamente a estocasticidade dos fenômenos estudados. Por enquanto, não há na teoria ligüística nenhuma

proposta que atribua significado a essa consolidação; a tentativa de adaptar um modelo de Gibbs pelo menos sugere perguntas a serem feitas. Por exemplo, que condições devem satisfazer as restrições definindo a função "energia", para assegurar a unicidade ou ao menos a cardinalidade baixa do conjunto de soluções.

No nossa pesquisa anterior com o modelo otimalista foi visto que a definição de restrições lingüisticas como propriedades de pés rítmicos consegue levar a um bom modelo de acentos secundários. Essa modelagem inicial admite extensões, de forma a incorporar outros fenômenos. O modelo matemático subjacente é o de busca de caminhos mínimos em um grafo, onde caminhos correspondem a subdivisões em pés, e os custos refletem violações de restrições. Na verdade, os custos não têm (ainda) significado lingüístico, e são um artifício usado na algoritmização do problema. Com os dados que foram testados até agora existe muita liberdade para a escolha de custos de maneira a modelar os dados reais. No próximo estágio da pesquisa espera-se obter um grande corpus de dados que restrinjam a escolha de custos de forma a se chegar a valores conceitualmente interpretáveis.

Para a adaptação de custos a grandes massas de dados será importante utilizar métodos de Otimização Inversa. Este assunto surgiu na última década; informalmente trata-se de fazer uma "conta de chegada"em problemas de otimização. Ou seja, sabendo-se a solução, como definir os custos de forma que essa solução seja ótima. Isso tem clara aplicação na modelagem a partir de dados reais. Existem já bons resultados que poderão ser aplicados a esta pesquisa, porém é altamente provável que seja necessário trabalho matemático no desenvolvimento de Otimização Inversa para problemas ainda não estudados na literatura.

Para validar estatisticamente os resultados de Sândalo et al (2003) precisamos considerar as sequências de sílabas codificadas utilizadas por Sotaq como realizações de uma cadeia de Markov de alcance variável. As probabilidades de transição dessa cadeia deveriam estar relacionadas à probabilidade de Gibbs subjacente ao modelo otimalista. Essa direção de pesquisa está atualmente na ordem do dia em nosso projeto. Um primeiro passo nessa direção foi sugerido pelo artigo Galves, Galves, Garcia e Peixoto (manuscrito 2003). A construção de procedimentos inferenciais rigorosos para essa classe de cadeias estocásticas serúm tema de pesquisa importante do presente projeto.

Essas considerações nos fazem re-encontrar num nível mais alto a idéia inicial que em 1993 deu origem ao grupo que apresenta este projeto. Probabilidades de Gibbs parecem ser bons candidatos

a modelos da interface sintaxe-ritmo. Nesse modelo o ritmo é codificado atraves de um potencial termodinâmico. Atualmente estamos em condições de modelar especificamente esse potencial. O objetivo do presente projeto é discutir as relações entre as propriedades deste potencial e os padrões rítmicos característicos das realizações da cadeia correspondente.

3.7 Laboratório de fonologia do ritmo

Nossa pesquisa matemática visa contituir um quadro conceitual rigoroso para o tratamento quantitativo de um conjunto de questões lingüísticas. Os modelos matemáticos fazem predições cuja realidade lingüística e exatidão devem ser comprovadas experimentalmente. Esse é o objetivo do laboratório de fonologia do ritmo que será instalado no NUMEC. Esse laboratório dá continuidade ao trabalho de contrução de ferramentas computacionais que foram desenvolvidas no quadro do Projeto Temático FAPESP *Padrões rítmicos, fixação de parâmetros e mudança lingüística* e do do Projeto CNPq (Edital 2000) *Técnicas probabilísticas para identificação de padroes rítmicos com aplicação à lingüística*.

A primeira ferramenta computacional para análise do sinal acústico de fala que nossa equipe desenvolveu é o programa *Vocale* que marca automaticamente as fronteiras dos intervalos consonantais e vocálicos (cf Garcia, Gut e Galves, 2002). O objetivo dessa marcação é a implementação automática da abordagem proposta por Ramus, Nespor e Mehler (1999). Esse programa está atualmente em fase β. Ele pode ser obtido livremente para uso acadêmico no endereço `http://www.ime.usp.br/~tycho/prosody`. O desenvolvimento ulterior do programa Vocale é uma das atividades previstas em nosso projeto.

Além do programa *Vocale*, nossa equipe também desenvolveu o programa *Piccolo* cujo objetivo é o cálculo automático da função sonoridade, seguindo a abordagem introduzida em Galves et al. (2002). Ele atua como um módulo do programa de uso público *Praat* que pode ser obtido no endereço `http://www.praat.or`. *Piccolo* também pode ser obtido livremente para uso acadêmico no endereço `http://www.ime.usp.br/~tycho/prosody`. Uma das atividades previstas em nosso projeto é comparar as medições feitas pelo programa *Piccolo*, a partir do espctrograma do sinál acústico de fala, com as medições da sonoridade obtidas diretamente com o uso de um laringógrafo.

Até aqui nossas medições utilizaram os corpora de língua constituídos pela equipe de Jacques Mehler no Laboratoire de Sciences Cognitives de Paris e corpora-pilotos de português brasileiro e europeu feitos por nossa equipe. No presente projeto vamos ampliar esse conjunto de dados, acrescentando novas amostras de português europeu e brasileiro, além de línguas indígenas brasileiras.

4 Resultados de auxílios anteriores

O presente projeto de pesquisa situa-se na continuação dos seguintes projetos:

- Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística*, processo 1998/03382-0;
- Projeto CNPq (Edital 2000) *Técnicas Probabilísticas de Identificação de Padrões Rítmicos com Aplicação à Lingüística*.

Apresentamos em anexo os relatórios finais dos referidos projetos. O relatório do projeto temático inclui um CD-Rom contendo cópia dos artigos matemáticos e linguísticos já produzidos. Esses artigos também podem ser obtidos a partir da página do projeto *Padrões rítmicos, fixação de parâmetros e mudança linguística* no endereço:

`http://www.ime.usp.br/~tycho/papers/index.html`.

4.1 Relatório Científico do Projeto: Técnicas Probabilísticas de Reconhecimento de Padrões, com Aplicações à Lingüística.

O projecto (CNPq 465928/2000-5) produziu os seguintes trabalhos

1. *Sonority as a basis for rhythmic class discrimination*, por A. Galves, J. Garcia, D. Duarte e Ch. Galves, em Proceedings of the Speech Prosody 2002 conference (ISBN 2-9518233-0-4), 323-326, 2002. Pode ser obtido em
   `http://www.lpl.univ-aix.fr/sp2002/pdf/galves-etal.pdf`
2. *Vocale-a semi-automatic annotation tool for prosody*, por J. Garcia, U. Gut e A. Galves, em Proceedings of the Speech Prosody 2002 conference (ISBN 2-9518233-0-4). Pode ser obtido em `http://www.lpl.univ-aix.fr/sp2002/pdf/garcia-gut-galves.pdf`;
3. *The Sotaq optimality based computer program and secondary stress in two varieties of Portuguese*, por M. B. Abaurre, C. Galves, A. Mandel e F. Sândalo. Pode ser obtido em `http://roa.rutgers.edu/view.php3?roa=463`;

4. *The statistical analysis of acoustic correlates of speech rhythm*, por D. Duarte, A. Galves, N. Garcia, e R. Maronna. Trabalho apresentado no ZiF, no quadro do Projeto *the Sciences of Complexity*
   (`http://www.uni-bielefeld.de/ZiF/complexity`). Pode ser obtido em
   `http://www.ime.usp.br/~tycho/tipal/results.html`

5. *Coupling, renewal and perfect simulation of chains of infinite order*, por R. Fernández, P. Ferrari e A. Galves, Notas para um minicurso na 5^a Escola Brasileira de Probabilidade. Pode ser obtido em
   `http://www.ime.usp.br/~galves/livro/revised2.pdf`.

O projeto também produziu os seguintes artigos que estão em fase final de redação

1. *Fluctuations of the empirical entropy of a chain of infinite order*, por D. Gabrielli, D. Guiol e A. Galves. Segue em anexo cópia de versão preliminar do artigo;

2. *Unbounded variable lenght Markov chains*, por D. Duarte, A. Galves, N. L. Garcia e C. Peixoto. Segue em anexo cópia de versão preliminar do artigo;

3. *Bootstrapping and Central Limit Theorem for Chains of Infinite Order*, por P. Collet, D. Duarte e A. Galves. Segue em anexo cópia de versão preliminar do artigo.

4. *Does Maurer's conjecture hold for mixing processes?*, por M. Abadi, P. Ferrari e A. Galves. Segue em anexo cópia de versão preliminar do artigo.

Além disso o projeto produziu o programa Vocale que segmenta automaticamente o sinal acústico de fala em intervalos vocálicos e consonantais. Este programa pode ser obtido no sítio `http://www.ime.usp.br/~tycho/prosody/vocale`.

Finalmente a equipe do projeto constituiu um corpus anotado de amostras de fala. Este corpus está aberto à pesquisa acadêmica e pode ser acessado no sítio `http://www.ime.usp.br/~tycho/prosody/data`.

O projeto foi muito bem sucedido. Matematicamente ele desenvolveu um quadro conceitual rigoroso para a modelagem de contornos rítmicos em línguas naturais. O melhor exemplo disso são as Notas escritas por Fernández, Ferrari e Galves. Estas ferramentas tem um interesse matemático autônomo, como é exemplificado pelos artigos matemáticos elencados neste relatório.

A utilização destas ferramentas matemáticas no problema da caraterização das classes rítmicas das línguas naturais foi extremamente bem sucedida. Um sinal claro disso é o fato do artigo *Sonority as a basis for rhythmic class discrimination*, de Galves, Garcia, Duarte e Galves, ter sido um dos quatro trabalhos selcionados para apresentação oral na sessão "Tipologia das línguas e classes rítmicas" do congresso Speech Prosody 2002, realizado em Aix-en-Provence. Este congresso é certamente o mais importante congresso na área de Fonética e Fonologia Laboratorial a ser realizado no ano de 2002.

Outro sinal do sucesso de projeto está no conjunto de citações de nossos trabalhos feitas no artigo de Franck Ramus em torno do qual a sessão "Tipologia das línguas e classes rítmicas" se organizava. O artigo de Franck Ramus se chama *Acoustic correlates of linguistic rhythm: perspectives* e pode ser obtido no endereço
`http://www.lpl.univ-aix.fr/sp2002/pdf/ramus.pdf`.

Também no congresso Speech Prosody 2002 tivemos aceito para apresentação em forma de poster o artigo em que apresentávamos o programa Vocale. Trata-se de um idéia original desenvolvida no quadro do projeto, com perspectivas extremamente interessantes de desenvolvimento. A versão atual funciona com boa precisão em amostras de Português Brasileiro e de Inglês. Estamos atualmente trabalhando numa versão com ajuste automático dos parâmetros, utilizando técnicas de *"boosting"*.

Finalmente quero mencionar que o sucesso do projeto se deve em boa parte à qualidade e à quantidade dos conjuntos de dados analisados. Trata-se de um trabalho de coleta de amostras, e em seguida de anotação manual extremante longo e que exige alto treinamento da equipe. Estamos atualmente, graças ao programa Vocale e à nova abordagem introduzida no artigo *Sonority as a basis for rhythmic class discrimination*, de Galves, Garcia, Duarte e Galves, numa nova etapa de anotação automática dos dados. Isso permitirá possivelmente em breve espaço de tempo uma consolidação dos critérios estatísticos e um refinamento dos modelos estocásticos dos contornos acentuais.

O conjunto de ponteiros dos trabalhos, programa Vocale e corpora mencionados neste relatório está resumido na página `http://www.ime.usp.br/~tycho/tipal/results.html`.

4.2 Resumo dos resultados do projeto temático: Padrões rítmicos, fixação de parâmetros e mudança lingüística

Ao longo dos quatro anos e meio do projeto, apresentamos 5 relatórios parciais com uma lista detalhada das diversas atividades realizadas pela equipe. O objetivo deste texto é fazer um balanço do caminho percorrido durante esse período, e apontar para as direções nas quais a pesquisa continua. Neste balanço, contemplaremos os três aspectos básicos do projeto: atividades de pesquisa propriamente dita, atividades de construção de corpora eletrônicos e ferramentas computacionais para a análise, e atividades de formação de recursos humanos. Finalmente, apresentaremos alguns indicadores do impacto do projeto na comunidade científica internacional.

4.2.1 Os avanços da pesquisa.

No sumário do nosso projeto, dizíamos que para desenvolver a pesquisa proposta, deveríamos:

1. Fazer uma descrição detalhada da mudança da sintaxe de colocação de clíticos em textos escritos por autores portugueses nascidos entre 1550 e 1850, descrevendo as gramáticas envolvidas. 2. Fazer uma descrição detalhada dos aspectos fonéticos relevantes para a identificação dos padrões rítmicos do Português Europeu Moderno e do Português Brasileiro. 3. Desenvolver um modelo linguístico-matemático para a noção de padrão rítmico. 4. Fazer um modelo formal de aquisição, relacionando prosódia e sintaxe na interface da gramática com o sistema Articulatório-Perceptual. 5. Desenvolver uma metodologia permitindo encontrar nos textos escritos os traços característicos do padrão rítmico da língua falada.

O primeiro ponto foi amplamente desenvolvido e levou a uma descrição detalhada da mudança da colocação de clíticos em textos de autores portugueses nascidos entre 1542 (Diogo do Couto) e 1836 (Ramalho Ortigão). Os dados nos quais se baseiam essa descrição estão coletados numa base de dados que contém 24.575 orações representando a totalidade das ocorrências de orações com clíticos em 20 textos do Corpus Tycho Brahe, num total de 941.031 palavras. Os resultados da pesquisa

com a descrição da evolução da colocação dos clíticos no período considerado se encontra sintetizada no artigo "The Change in Clitic Placement from Classical to Modern European Portuguese: Results from the Tycho Brahe Corpus"(Galves, Britto e Paixão de Sousa 2003). Nesse trabalho se encontra uma ampla lista de referências aos diversos trabalhos sobre o assunto produzidos pela equipe do projeto. Essa pesquisa lançou uma nova luz sobre a história do Português, trazendo claras evidências de que a mudança na colocação de clíticos se deu na virada do século 17 para o século 18, e não no início do século 17, como defendido por Martins (1994), ou na segunda metade do século 18, como proposto na nossa hipótese inicial. O rico material colhido ainda está sendo aproveitado para vários outros enfoques, complementares com a análise da colocação de clíticos, que vão constituindo uma verdadeira radiografia da língua portuguesa num período até agora muito pouco conhecido desse ponto de vista.

Em relação à questão dos padrões rítmicos colocada no ponto 2., houve dois encaminhamentos paralelos potencialmente convergentes. O primeiro deles redundou numa abordagem otimalista da atribuição de acentos secundários e redução vocálica no PE e PB. Os resultados dessa pesquisa são apresentados no artigo “The Sotaq optimality based computer program and secondary stress in two varieties of Portuguese” (Abaurre, Galves, Mandel e Sândalo 2002). Um desdobramento computacional dessa atividade foi a elaboração do programa Sotaq que será apresentado na segunda parte desta síntese. A segunda vertente trabalhou com a noção de correlatos acústicos do ritmo, seguindo a pista sugerida por um conjunto de pesquisas recentes na área iniciadas por Ramus, Nespor e Mehler (1999). Esse esforço levou a propor uma medida de sonoridade como base para distinguir as classes rítmicas. Esse ponto de vista se encontra apresentado no artigo "Sonority as a basis for rhythmic class discrimination"(Galves, Garcia, Duarte e Galves, 2002). Um desdobramento computacional dessa atividade foi a elaboração dos programas Piccolo e Vocale que serão apresentados na segunda parte desta síntese. Essas duas vertentes, apesar de fundadas em metodologias e quadros teóricos distintos convergiram para a mesma conclusão, já proposta na literatura anterior, mas só agora consolidada em fatos e análises mais sistemáticas: o PE e o PB instanciam ritmos de natureza fundamentalmente diferente. O trabalho desenvolvido contribuiu substancialmente em caracterizar cada um desses ritmos.

Em relação ao terceiro ponto, o modelo lingüístico- matemático da noção de padrão rítmico, as

duas vertentes acima descritas redundaram em propostas potencialmente convergentes. O modelo otimalista dá evidências de que as unidades rítmicas são constituídas diferentemente em Português Europeu e em Português Brasileiro. Como conseqüência, sugere que o processo estocástico constituído pelas palavras sucessivas de um texto codificadas segundo a posição dos seus acentos principais, número de sílabas, posição em relação aos sintagmas fonológicos e palavras prosódicas às quais pertencem, tem leis diferentes. Mais precisamente, sugere que se modelarmos este processo através de uma cadeia de Markov de alcance variável, as funções contexto correspondentes ao PE e ao PB são distintas: no caso do primeiro, elas esquecem o passado sempre que encontram uma fronteira de palavra prosódica, e no segundo, sempre que encontram uma fronteira de sintagma fonológico. Isso permitirá potencialmente a identificação das características rítmicas subjacentes aos textos históricos do Corpus Tycho Brahe. Esse trabalho está em andamento. A segunda vertente apresentada no item 2. também propõe uma resposta para o item 3. Com efeito, há evidências empíricas de que a sonoridade pode ser bem modelada por uma cadeia quantificada com dois estados subjacentes. Essas cadeias têm a característica seguinte. As cadeias discretas subjacentes à sonoridade de cada língua têm leis que diferem de língua para língua. No entanto, as duas distribuições correspondentes aos dois estados subjacentes são universais. Isso permite identificar estatísticamente o ponto de corte separando as duas zonas de sonoridade, e em seguida, codificar a sonoridade através de uma cadeia de ordem infinita assumindo dois valores. São essas cadeias que carregariam todas as informações rítmicas de cada língua. A metodologia estatística necessária para levar à frente este programa está atualmente sendo desenvolvida. Os primeiros resultados estão na tese de Denise Duarte defendida em 2003, "Aproximações markovianas e reamostragem em cadeias de ordem infinita com aplicação à linguística", e nos artigos "Markov approximations and the bootstrap for chains of infinite order"(Collet, Duarte e Galves, em andamento), e "Stochastic modelling of the speech sonority: quantization and cross language estimation of the Cut Point"(Cassandro, Collet, Duarte, Galves e Garcia, em andamento). Uma versão preliminar resumida do segundo artigo se encontra no texto "An universal linear relation among acoustic correlates of rhythm". O quarto ponto está na origem mesmo deste projeto temático. O modelo proposto se encontra resumido nos artigos "A Statistical-Physics approach to language acquisition and language change"(Cassandro, Collet, Galves e Galves 1999) e “Identifying features in the presence of competing evidence, the case of first language

acquisition" (Fernández e Galves 2000). O modelo matemático apresentado nesses artigos é a base de toda a modelagem estocástica dos dados do projeto. O grande desafio, que foi iniciado neste projeto mas deverá ser colocado como cerne do próximo, é alimentar e validar empiricamente esse modelo com os dados quantitativos que têm sido produzidos nas diversas linhas de pesquisa que compuseram este temático. A tarefa de obtenção de grandes quantidades de dados devidamente categorizados foi a prioridade do período que se encerra agora. Além dos dados em si, isso produziu metodologias e ferramentas de grande valia para futuras pesquisas. Esses dados constituirão a base da próxima fase do projeto proporcionando condições para que a modelagem atinja todo o seu potencial explicativo.

O trabalho sobre o quinto ponto, que constituí um dos aspectos mais originais, e certamente o mais ambicioso do projeto, deu resultados preliminares auspiciosos, apresentados no relatório "Modelagem de contornos acentuais do Português através de cadeias de Markov de alcance variável"no artigo "Correlates of rhythm in written texts of Brazilian and Modern European Portuguese"(A. Galves, C. Galves, N. Garcia e C. Peixoto em andamento - a ser submetido à revista JASA). O avanço mais espectacular é o desenvolvimento de uma metodologia de identificação de padrões rítmicos nos textos escritos que articula a noção probabilística de Cadeias de Markov de alcance variável, com os resultados recentes de várias pesquisas sobre as diferenças prosódicas do PE e do PB, várias delas produzidos no próprio projeto. No estado atual da pesquisa, ainda não foi possível confirmar uma das hipóteses centrais do projeto, a saber que o ritmo da língua muda antes da sintaxe, o que suportaria a hipótese de que a mudança rítmica desencadeou a mudança sintática. Com a metodologia desenvolvida, obtém-se uma clara separação do PE e do PB, o que constituí um avanço importante, mas é impossível detectar nos autores do Corpus Tycho Brahe um ponto em que o ritmo muda (segundo a nossa hipótese, de um ritmo idêntico ao PB para o ritmo do PE). Os autores do Corpus Tycho Brahe ao contrário do esperado, mantêm um comportamento bastante homogêneo. Isso não significa que devamos abandonar a nossa hipótese, uma vez que há índices fortes da existência de uma mudança acontecida na prosódia da língua portuguesa entre o século 16 e o século 18. Mostra que não dispomos ainda das ferramentas necessárias, ou da plena capacidade para usá-las e interpretá-las. Traços extremamente interessantes emergem porém do que já obtivemos, apontando para o fato de que, diferentemente da nossa hipótese inicial, os padrões rítmicos não podem ser detectados independentemente mas vêm "borrados" por outros fenômenos, sintáticos

e estilísticos. A continuação desta pesquisa, o refinamento das ferramentas e das interpretações, continuam fortemente na ordem do dia.

4.2.2 Corpora e ferramentas computacionais

O principal produto eletrônico do projeto é o Corpus Anotado do Português Histórico Tycho Brahe, livremente acessível à comunidade acadêmica através do endereço http://www.ime.usp.br/ tycho/corpus. Nossa previsão inicial era que tivesse 1.000.000 de palavras. Essa previsão foi ultrapassada, já que totaliza 1.851.619 palavras, das quais 1.019.191 já se encontram em versão morfologicamente etiquetada. A construção do Corpus envolveu a elaboração de ferramentas computacionais de anotação, das quais as principais são o etiquetador automático desenvolvido por Marcelo Finger, e o analisador automático para o Português obtido a partir do treinamento de um analisador universal desenvolvido na Universidade de Pensilvânia por Dan Bickel. O treinamento desse analisador foi possível graças ao desenvolvimento de um sistema de anotação sintática, nos moldes do sistema proposto por Taylor e Kroch (1998), para o Inglês Médio, e da anotação manual, conforme esse sistema, de um texto de 50 000 palavras por Helena Britto, pós-doutoranda do projeto. O estudo dos padrões rítmicos envolveu a elaboração de vários programas computacionais de grande potencialidade para a pesquisa do ritmo da fala em geral. O programa Sotaq foi desenvolvido por Arnaldo Mandel, retomando e aperfeiçoando um protótipo feito inicialmente por Pierre Collet e Antonio Galves. Maiores detalhes sobre o programa Sotaq, sobre o modelo otimalista que estamos utilizando e sobre a própria Teoria da Otimalidade podem ser encontrados na página http://www.ime.usp.br/ tycho/prosody/sotaq . Dois outros programas foram desenvolvidos no âmbito da pesquisa sobre as classes rítmicas. Vocale http://www.ime.usp.br/ tycho/prosody/vocale é uma ferramenta para a anotação automática de intervalos vocálicos e consonantais que toma como input os arquivos sonoros sem nenhuma anotação manual. Piccolo, http://www.ime.usp.br/ tycho/prosody/sonority , é uma versão mais simples de Vocale que permite medir a sonoridade de intervalos sucessivos de fala.

4.2.3 Formação de recursos humanos

Durante a vigência do auxílio, o projeto produziu

- 6 doutorados, dos quais 3 já defendidos - 10 mestrados, dos quais 4 já defendidos - 20 projetos

de Iniciação Científica dos quais 14 já concluídos - 7 projetos de Treinamento Técnico - 4 projetos de Pós-Doutorado.

4.2.4 Índices de impacto e desdobramentos.

O impacto do Corpus Tycho Brahe e da metodologia de anotação usada na sua elaboração pode ser medida por vários indicadores: - até o presente dia, cerca de 300 pesquisadores do mundo inteiro pediram senha para acessar os textos (cf. lista em anexo) - os sistemas de anotação morfológica e sintática que desenvolvemos foram adotados pelo Projeto português de Corpus Dialetal Sintático (Cordial Sin), coordenado por Ana Maria Martins no Centro de Lingüística da Universidade de Lisboa. Uma bolsista do CordialSin visitou o nosso projeto durante o mês de janeiro de 2000 para aprender a usar o nosso sistema de anotação morfológica e nossas ferramentas de correção. Em maio de 2002, Helena Britto foi convidada a passar 1 mês em Lisboa para apresentar e discutir o sistema sintático que ela desenvolveu no âmbito do projeto. - O reconhecimento do interesse de trabalhar com grandes corpora anotados vem crescendo no Brasil, a partir da divulgação do nosso trabalho. Temos interagido com vários grupos de pesquisa que trabalham com a descrição diacrônica e sincrônica do português do Brasil, com vistas a futuras parcerias e trocas. - Fomos convidados juntos com outros projetos de vários países europeus para entrar como colaboradores externos num projeto canadense de Corpus anotado de textos franceses do séc. 9 ao séc. 17 (cf. carta em anexo do Prof. Paul Hirschbuhler da Universidade de Ottawa).

O sucesso da pesquisa em modelagem estocástica da fala pode ser medida pelos seguintes fatos: - António Galves foi convidado para apresentar os aspectos matemáticos do projeto como conferência plenária do mais importante congresso internacional da área de Física-Estatística (StatPhys 1998, Paris). O texto da conferência foi publicado no artigo "A Statistical-Physics approach to language acquisition and language change", na prestigiosa revista Physica. - O projeto Técnicas probabilísticas de identificação de padrões com aplicações à lingüística (TIPAL), oriundo deste projeto, foi contemplado no nível mais alto de financiamento, aproximadamente 100.000 reais, no Edital 2000 do CNPq. Este foi o único projeto contemplado nesse nível nas áreas de Matemática/Estatística e Lingüística. - O artigo "Sonority as a basis for rhythmic class discrimination", foi escolhido num dos Congresso internacionais mais importantes da área Speech Prosody 2002, como uma das 4 comunicações

selecionadas para a sessão plenária "Prosody and Linguistic Typology".

Enfim, o impacto e o reconhecimento da validade da nossa proposta interdisciplinar podem ser avaliados a partir dos seguintes fatos: - O Instituto do Milênio para o Avanço Global Integrado da Matemática no Brasil incorporou o projeto como uma das suas áreas de atuação. - Fomos convidados para organizar sessões de trabalho de um mês cada uma, no Complexo Inter- disciplinar da Universidade de Lisboa (fevereiro de 2000), e no Zentrum fur Interdisziplinare Forschung - ZiF- , da Universidade de Bielefeld (julho de 2001), no âmbito do Ano da Complexidade. Em ambos os casos, os eventos foram amplamente financiados com recursos das instituições hospedeiras, com auxílios complementares de outras instituições europeias e americanas. - O Projeto está na base da criação do "Núcleo de Modelagem Estocástica e Complexidade"(NUMEC) da USP, já constituído, e do "Núcleo de Estatística e Identificação de Padrões em Grandes Corpora de Língua"(NEIPACL), em fase de implantação, na UNICAMP.

5 Apresentação da Equipe

5.1 Pesquisadores principais

1. Jefferson Antonio Galves (Estatística, USP)

   Coordenador geral do projeto.

   Probabilista. Pesquisador 1A do CNPq. Membro titular da Academia Brasileira de Ciências e Comendador da Ordem Nacional do Mérito Científico. Foi coordenador do Núcleo de Excelência *Fenômenos Críticos em Probabilidades e Processos Estocásticos* e do projeto CNPq (Edital 2000)*Técnicas Probabilísticas de Identificação de Padrões Rítmicos com Aplicação à Lingüística*- TIPAL- , bem como pesquisador principal do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística*. É coordenador do *Núcleo de pesquisa em Modelagem estocástica e complexidade* da Universidade de São Paulo (NUMEC).

   Seus interesses de pesquisa cronologicamente foram: os sistemas markovianos com muitos componentes, o comportamento hidrodinâmico e metaestável de sistemas estocásticos, as aproximacões exponenciais de cadeias de ordem infinita, e finalmente, tem se interessado por questões de inferência estatística em cadeias de ordem infinita. Desde 1993 vem utilizando esses conceitos na modelagem da aquisição e da mudança linguística, e na caracterização de padrões rítmicos nas línguas naturais.

2. Maria Bernadete Marques Abaurre (Linguística, UNICAMP)

   Coordenará a análise fonológica.

   Fonóloga. Pesquisadora 1A do CNPq. Foi pesquisadora principal do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística*. Foi coordenadora do Grupo de Trabalho de Fonética/Fonologia do Projeto temático *Gramática do Português falado*. Coordena o Projeto integrado *A relevância teórica dos dados singulares na aquisição da linguagem escrita*. É membro do projeto FAEP/UNICAMP *Fonologia prosódica do português: estudo acústico e modelagem a partir da teoria da otimalidade*. Seus principais interesses de pesquisa são a modelagem dos padrões rítmicos das línguas naturais, a interface sintaxe/fonologia e a aquisição

da escrita.

3. Charlotte Galves (Linguística, UNICAMP)

   Coordenará da pesquisa sobre a modelagem estocástica do ritmo em textos escritos.

   Sintaticista. Pesquisadora 1B do CNPq. Foi coordenadora do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística* e membro do Projeto TIPAL. Coordena atualmente o Projeto CAPES-DAAD *Mineração de dados em grandes corpora de língua*. É membro dos projetos FAEP/UNICAMP *Fonologia prosódica do português: estudo acústico e modelagem a partir da teoria da otimalidade* e *Modelagem estocástica da fala*. É membro do Comitê gestor do *Núcleo de pesquisa em Modelagem estocástica e complexidade* da Universidade de São Paulo (NUMEC). Seus principais interesses de pesquisa têm sido a comparação do português europeu e brasileiro, a história da sintaxe do português baseada em grandes corpora e o papel do ritmo na mudança gramatical.

4. Nancy Lopes Garcia (Estatística, UNICAMP)

   Coordenará a pesquisa sobre inferência em cadeias de ordem infinita com aplicações à linguística.

   Probabilista. Pesquisadora 2A do CNPq. Foi membro do Núcleo de Excelência *Fenômenos Críticos em Probabilidades e Processos Estocásticos*, do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística* e do Projeto TIPAL. Coordena o Projeto FAEP/Unicamp *Modelagem estocástica da fala*. É membro do projeto FAEP/UNICAMP *Fonologia prosódica do português: estudo acústico e modelagem a partir da teoria da otimalidade*. Seus principais interesses de pesquisa são processos pontuais, processos estocásticos com aplicação à linguística, cadeias de ordem infinita, e simulação.

5. Arnaldo Mandel (Computação, USP)

   Coordenará a pesquisa sobre desenvolvimento de algoritmos e otimização aplicados à identificação de padrões rítmicos.

   Cientista da Computação. Responsável pelo desenvolvimento da rede do IME/USP. Possivelmente um dos maiores especialistas em construção e gerenciamento de redes UNIX no Brasil.

Foi membro do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística* e do Projeto TIPAL. É membro do Comitê gestor do *Núcleo de pesquisa em Modelagem estocástica e complexidade* da Universidade de São Paulo (NUMEC). É autor do programa *Sotaq*.

5.2 Pesquisadores colaboradores

- dentro do Estado

1. Miguel Abadi (Pós-doutorado/FAPESP, IME/USP)

   Probabilista. Doutorou-se com Antonio Galves. Seu trabalho de pesquisa é sobre aproximação exponencial para as leis dos instantes de ocorrência de eventos raros em cadeias de ordem infinita com boas propriedades de mistura. Atualmente está usando esses resultados na obtenção de estimadores da entropia de cadeias de ordem infinita, e no refinamento de resultados prévios da equipe referentes ao teorema limite central de reamostragem em cadeias de ordem infinita.

2. Anne Cros (Pós-doutorado/CNPq, IME/USP)

   Física. Recém-doutorou-se pela Universidade Aix-Marseille. O assunto de sua tese é a identificação de padrões estatísticos no caminho para a turbulência de um fluído. No projeto, ela trabalhará na modelagem estocástica da sonoridade e outros correlatos acústicos do ritmo.

3. Didier Demolin (Universidade Livre de Bruxelas/Linguística, USP)

   Fonólogo experimental. Ensina a fonética, fonologia e etnomusicologia na Universidade Livre de Bruxelas, onde criou e dirige o Laboratório de Fonologia. Atua como Pesquisador Visitante na Universidade de São Paulo desde agosto de 2003. Fez trabalho de campo em línguas da África e do Brasil. Suas áreas de interesse são a Fonologia de Laboratório, a fisiologia da fala, e a evolução da linguagem. No projeto ele coordenará as atividades de fonologia laboratorial, trabalhando particularmente na busca das evidências empíricas que corroboram as predições dos modelos matemáticos.

4. Ronaldo Dias (Estatística, UNICAMP)

Estatístico. Pesquisador 2B do CNPq. Coordena o projeto CNPq (edital 2003) *Testes não paramétricos para proximidade de duas distribuições*. Seus principais interesses de pesquisa são modelos estatísticos não paramêtricos e métodos computacionais.

5. Jesús Enrique García (Estatística, UNICAMP)

   Probabilista. Participou como estudante de doutorado do Núcleo de Excelência *Fenômenos Críticos em Probabilidades e Processos Estocásticos*. Foi membro do projeto TIPAL e do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística*. É o principal desenvolvedor dos programas *Vocale* e *Piccolo*. Seu principal tema de pesquisa atualmente é a modelagem estocástica da sonoridade da fala.

6. Verónica González-López (Estatística, UNICAMP)

   Estatística. Obteve seu Ph. D. em (Estatística) na Universidade de São Paulo. Fez pós-doutorados no Departamento de Matemática da UBA, com auxílio da UBATEC /Argentina - orientado por Victor Yohai em Estimação Robusta, e no Instituto de Matemática da USP, com apoio da FAPESP. No projeto sua principal atividade de pesquisa será a modelagem por cópulas de correlatos acústicos do ritmo na fala.

7. Cláudia Monteiro Peixoto (Estatística, USP)

   Probabilista. Pesquisadora 2C do CNPq. Foi membro do Núcleo de Excelência *Fenômenos Críticos em Probabilidades e Processos Estocásticos*, bem como do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística*. Seus principais interesses de pesquisa são processos estocásticos e modelagem probabilística do ritmo em textos escritos.

8. Filomena Spatti Sândalo (Linguística, UNICAMP)

   Fonóloga e sintaticista. Fez pós-doutorado no Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística*. Coordena o projeto FAEP/UNICAMP *Fonologia prosódica do português: estudo acústico e modelagem a partir da teoria da otimalidade*. É membro do projeto *Endangered Languages Documentation Project*, da SOAS (School of Oriental and African Languages), Inglaterra. Seus principais interesses de pesquisa são a modelagem dos pa-

drões rítmicos das línguas naturais, a interface sintaxe/fonologia, a caracterização dos domínios prosódicos nas línguas naturais e o estudo de línguas indígenas brasileiras .

9. Luciana Storto (Linguística, USP)

   Fonóloga e sintaticista. Pesquisadora 2C do CNPq. Trabalha com a descrição da língua Karitiana, da familia Arikem, tronco Tupi, desde 1992. Seus interesses de pesquisa são tom, acento, padrões rítmicos, especialmente no que diz respeito à interface entre fonologia e sintaxe.

5.3 Fora do Estado de São Paulo

1. Denise Duarte (Matemática, UFGo)

   Estatística. Realizou sua tese de doutorado no âmbito do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística* e do Projeto TIPAL. Sua pesquisa atual está ligada à modelagem estocástica da sonoridade nas línguas naturais e ao desenvolvimento do teorema-limite central para estimadores re-amostrados de cadeias de ordem infinita com aplicações às cadeias subjacentes à sonoridade.

5.4 Colaboradores externos

1. Marzio Cassandro (Física, La Sapienza, Roma)

   Físico teórico. Foi colaborador do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística* e do Projeto TIPAL. Participou da sessão de trabalho *Rhythmic patterns, parameter setting and language change*, organizada no âmbito do ano da Complexidade no ZiF*.

2. Pierre Collet (CNRS, Paris)

   Matemático. Foi colaborador do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística* e do Projeto TIPAL. Participou das sessões de trabalho *Statistical Physics, Pattern Identification and Language Change* em Lisboa** e *Rhythmic patterns, parameter setting and language change* no ZiF*.

3. Emmanuel Dupoux (Laboratoire de Sciences Cognitives et Psycholinguistiques, École des Hautes Études en Sciences Sociales e École Normale Supérieure, Paris)

   Psicolinguista. Diretor do Laboratoire de Sciences Cognitives et Psycholinguistique. Participou das sessões de trabalho *Statistical Physics, Pattern Identification and Language Change* em Lisboa** e *Rhythmic patterns, parameter setting and language change no ZiF**.

4. Roberto Fernández (Laboratoire de Mathématique Raphael Salem, Rouen)

   Físico matemático. Foi colaborador do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística* e do Projeto TIPAL. Participou das sessões de trabalho em Lisboa** e no ZiF*.

5. Sónia Frota (Linguística, Lisboa)

   Foi uma das colaboradoras do Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística*, onde realizou um trabalho importante sobre a comparação do ritmo em Português Europeu e Brasileiro. Foi uma das organizadoras da sessão *Statistical Physics, Pattern Identification and Language Change* em Lisboa**. Participou da sessão de trabalho *Rhythmic patterns, parameter setting and language change* no ZiF.

6. Ulrike Gut (Linguística, Freiburg)

   Fonóloga. É co-autora do Programa *Vocale*. Participou da sessão de trabalho *Rhythmic patterns, parameter setting and language change* no ZiF. É membro do projeto CAPES-DAAD *Mineração de dados em grandes corpora de língua* que foi elaborado durante essa sessão.

7. Anthony Kroch (Linguística, Universidade da Pensilvânia)

   Sintaticista, um dos mais importantes pesquisadores da atualidade em linguística histórica, onde ele introduziu novos métodos estatísticos para a análise dos dados. Teve um papel determinante na elaboração do Corpus histórico do Português Tycho Brahe, e na discussão da pesquisa sobre a modelagem da interação sintaxe-ritmo desenvolvida no Projeto Temático *Padrões rítmicos, fixação de parâmetros e mudança linguística*. Participou da sessão de trabalho

Rhythmic patterns, parameter setting and language change no ZiF* e de vários encontros do projeto na UNICAMP.

8. Ricardo Lima (CNRS, Marseille)

   Físico. Participou das sessões de trabalho *Statistical Physics, Pattern Identification and Language Change***, e *Rhythmic patterns, parameter setting and language change* no ZiF*.

9. Marina Nespor (Linguística, Ferrara)

   Uma das importantes fonólogas da atualidade. Participou das sessões de trabalho *Statistical Physics, Pattern Identification and Language Change* em Lisboa**, e *Rhythmic patterns, parameter setting and language change* no ZiF*.

10. Sharon Pepperkamp (Laboratoire de Sciences Cognitives et Psycholinguistiques, École des Hautes Études en Sciences Sociales e École Normale Supérieure, Paris)

    Fonóloga do Laboratoire de Sciences Cognitives et Psycholinguistique. Tem atuado juntamente com Emmanuel Dupoux como consultora do projeto. É coordenadora do projeto *Acquisition phonologique précoce : algorithmes et simulations*, do qual Antonio Galves é consultor.

11. Janet Pierrehumbert (Linguística, Northwestern University)

    Uma das mais importantes foneticistas da atualidade. Lançou a chamada Fonologia laboratorial, que renovou profundamente tanto a fonologia quanto a fonética, restabelecendo laços fortes entre esses dois aspectos da pesquisa sobre a fala. Iniciou a linha de pesquisa em fonologia probabilística. É uma das fontes de referência do nosso projeto. Desde 2001, tem acompanhado e discutido conosco a nossa pesquisa sobre sonoridade e ritmo. Coordena o projeto *Dynamics and Metastability in phonological grammar*, apoiado pela Fundação James Macdonnell.

12. Marina Vigário (Linguística, Universidade do Minho)

    Fonóloga. Participou das sessõe s de trabalho *Statistical Physics, Pattern Identification and Language Change* em Lisboa**, e *Rhythmic patterns, parameter setting and language change* no ZiF*. Sua tese sobre a fonologia da palavra em português, elaborada durante a vigência do

projeto *Padrões rítmicos, fixação de parâmetros e mudança linguística* tem grande relevância para a pesquisa desenvolvida no presente projeto.

* cf. http://www.physik.uni-bielefeld.de/complexity/index.html

** cf. http://www.ime.usp.br/ tycho/meetings/

Apresentamos em pasta anexa os Cadastros, Súmulas Curriculares e Curricula Lattes de todos os pesquisadores principais deste projeto. Apresentamos os Cadastros de todos os pesquisadores colaboradores do projeto atuando em instituições brasileiras.

6 Justificativa Orçamentária

6.1 Adaptação do NUMEC ao projeto

O projeto prevê um Laboratório de Fonética, e para uma adequada instalação será necessário:

1. cabine acústica
2. instalação elétrica especial
3. bancadas para aparelhos
4. mobiliário - mesas, cadeias, bancos, armários
5. ar-condicionado

Além do laboratório de fonética teremos os laboratórios de informática, para os quais será necessário:

1. mobiliário adequado - mesas, cadeiras, armários.
2. ar-condicionado

O NUMEC encontra-se com a estrutura pronta para receber o grupo de pesquisadores mas ainda há necessidade de alguns acabamentos: fiação, cabeamentos, vidros temperados, pisos adequados.

6.2 Laboratório de Fonética

Demolin (2001) defende a tese de que independentemente do modelo de fonologia adotado num projeto, a teoria fonológica deve basear-se em modelos que incorporem parâmetros dos subsistemas envolvidos na comunicação oral. Entre estes estão os princí pios que relacionam articulação no trato vocal e saída acústica, princípios aerodinâmicos, e princípios que expliquem como nosso sistema auditivo extrai informação do sinal acústico. Assim, a teoria fonológica deve incorporar os fatos estabelecidos pelos modelos de produção e percepção da fala. O laboratório de fonética proposto neste projeto vem de encontro a estes objetivos.

O equipamento solicitado resume-se, em grande parte, em computadores e programas, ligados a algumas peças de hardware. Um uso importante a ser feito dos computadores no Laboratório de Fonética é a análise acústica, através dos softwares livres Formants e Signal Explorer desenvolvidos no Laboratório de Fonologia da Universidade Livre de Bruxelas, além do hardware e software chamado PCQuirer, que permite também análises aerodinâmicas. Os primeiros geram análises acústicas de sons digitalizados em temos de freqüência, amplitude e duração, o que possibilita a identificação precisa da qualidade de todas as vogais e da maior parte das consoantes. Para se estudar nasalidade tanto em vogais como em consoantes, é preciso de medidas aerodinâmicas, utilizando máscaras específicas (orais e nasais) ligadas ao PCQuirer. Quando se trata de línguas pouco conhecidas, é importante que análises acústicas deste tipo sejam incluídas nos artigos científicos produzidos pelos pesquisadores associados ao laboratório a fim de dar respaldo ao trabalho descritivo. Os sons implosivos, por exemplo, só podem ser identificados através de medidas de pressão oral, que o PCQuirer também é capaz de fazer. Um dos objetivos em se atrelar o PCQuirer aos softwares de análise acústica é ter dados acústicos e aerodinâmicos de boa qualidade gravados simultaneamente. Assim, para cada dado, temos não apenas a sua freqüência, amplitude e duração, mas também a pressão e o fluxo de ar nas cavidades oral e nasal. Uma vantagem do equipamento acima citado é o fato dele ser portátil, ou seja, pode-se levá-lo até as aldeias indígenas para a elicitação de dados.

O equipamento chamado EGG (Sistema Electroglotográfico) é um laringógrafo, indispensável ao laboratório de Fonética a ser criado através desse projeto, pois quando se trabalha com tom ou tipos de fonação (qualidade de voz) apenas ele é capaz de produzir certas medidas representativas dos estados da glote e da laringe.

O melhor equipamento para gravar dados é o sistema Nagra V com discos rígidos. Ter este equipamento nos permitirá evitar perda de tempo digitalizando dados de DAT para computadores.

6.3 Equipamentos de Informática

Os dados coletados pelo projeto serão arquivos de som digitalizado e ocuparão vários GB. O processamento desses arquivos para análise pelas ferramentas utilizadas gera arquivos pelo menos 10 vezes maiores. Assim, o projeto requer um grande servidor de arquivos, bem como grande capacidade

computacional.

Dada a distribuição geográfica dos participantes do projeto, é conveniente haver uma certa descentralizaçào do armazenamento e processamento. Isso e mais o desejo de redundância, para confiabilidade, orienta a arquitetura de rede proposta, baseada em três servidores:

O servidor A será o principal depósito e servidor de arquivos do projeto. Será integrado à rede do IME, e servirá arquivos também ao resto do Instituto; sua administração será integrada à da rede do IME, através de uma chave kvm, donde se dispensam equipamentos de interface humana.

O servidor B, localizado no IEL, será um espelho dos arquivos do projeto, e proverá alguns serviços de processamento extra. Boa parte do software utilizado ou desenvolvido pelo projeto funciona no sistema GNU/Linux. Os pesquisadores do IEL têm mais familiaridade com o uso de windows, que preferem ter nos seus equipamentos de uso diário. Este servidor proverá o ambiente GNU/Linux para utilização daquele software por esses pesquisadores.

O servidor C, localizado no IMECC, destina-se a tarefas pesadas de processamento, principalmente o envolvendo tratamento estatístico de sinais.

O espelhamento do servidor de disco substituirá o uso de backup em fitas, que é muito custoso para a escala de dados que serão tratados.

Os equipamentos de mesa e notebooks serão utilizados pelos perquisadores para as tarefas usuais de processamento e comunicação. Como boa parte do trabalho será feito no campo, há necessidade de equipamentos portáteis.

7 Cronograma de Atividades

Nos próximos tres anos esforços de pesquisa serão concentrados nas seguintes tarefas:

7.1 Primeiro ano

1. Desenvolvimento de pesquisas dirigidas à demonstração de resultados assintóticos do tipo teorema-limite central para a reamostragem sequencial em cadeias de ordem infinita.

2. Desenvolvimento de pesquisa sobre inferência em cadeias de Markov de alcance variável, com aplicações na identificação de padrões rítmicos nas línguas naturais.

3. Modelagem estocástica da sonoridade e outros correlatos do ritmo.

4. Desenvolvimento de uma teoria de dualidade para o modelo otimalista que permita uma boa solução dos problemas inversos de otimização suscitados pela análise de padrões rítmicos em línguas naturais.

5. Estudo das propriedades dos estimadores da entropia em cadeias de ordem infinita.

6. Desenvolvimento de um modelo probabilístico para as classes rí tmicas em línguas naturais através de um funcional do tipo *energia livre*.

7. Desenvolvimento de uma distribuição pública do programa *Piccolo*

8. Desenvolvimento de um ambiente para preparação de dados para o programa *Sotaq*

9. Incorporação ao programa *Sotaq* de novos módulos lingüísticos.

10. Preparação de uma distribuição pública dos atuais corpora anotados de fala do projeto.

11. Desenvolvimento de programa estatístico para identificação de relações de dependência e identificação de cópulas em dados bi-dimensionais.

12. Realização de um trabalho contínuo de fonologia laboratorial para medições de correlatos acústicos do ritmo, e verificação empírica das predições dos modelos matemáticos.

7.2 Segundo ano

1. Continuação dos esforços de pesquisa dirigidas à demonstração de resultados assintóticos para a reamostragem sequencial em cadeias de ordem infinita.
2. Continuação do esforço de pesquisa sobre inferência em cadeias de Markov de alcance variável, com aplicações na identificação de padrões rítmicos nas línguas naturais.
3. Continuação do esforço de modelagem estocástica da sonoridade e outros correlatos do ritmo.
4. Continuação do desenvolvimento em um modelo probabilístico para as classes rítmicas em línguas naturais através de um funcional do tipo *energia livre*.
5. Incorporação ao programa *Sotaq* de novos módulos lingüísticos.
6. Extensão do programa estatístico para identificação de relações de dependência e identificação de cópulas para o caso de dados tri-dimensionais.
7. Desenvolvimento de uma versão do programa *Vocale* incorporando técnicas de *boosting* para o balanceamento dos classificadores.
8. Preparação de distribuição pública de novos corpora anotados de fala de português brasileiro e europeu.
9. Realização de trabalho de campo para coleta de amostras de fala de karitiana
10. Realização de um trabalho contínuo de fonologia laboratorial para medições de correlatos acústicos do ritmo, e verificação empírica das predições dos modelos matemáticos.

7.3 Terceiro ano

1. Continuação dos esforços de pesquisa dirigidas à demonstração de resultados assintóticos para a reamostragem sequencial em cadeias de ordem infinita.

2. Continuação do esforço de pesquisa sobre inferência em cadeias de Markov de alcance variável, com aplicações na identificação de padrões rítmicos nas línguas naturais.

3. Continuação do esforço de modelagem estocástica da sonoridade e outros correlatos do ritmo.

4. Continuação do desenvolvimento em um modelo probabilístico para as classes rítmicas em línguas naturais através de um funcional do tipo *energia livre*.

5. Incorporação ao programa *Sotaq* de novos módulos lingüísticos.

6. Desenvolvimento de uma distribuição pública do programa *Vocale*

7. Preparação de distribuição pública de novos corpora anotados de fala de português brasileiro e europeu.

8. Realização de trabalho de campo para coleta de amostras de fala de karitiana

9. Preparação de distribuição pública de corpus anotado de fala de karitiana.

10. Realização de um trabalho contínuo de fonologia laboratorial para medições de correlatos acústicos do ritmo, e verificação empírica das predições dos modelos matemáticos.

Estão previstos três encontros de trabalho, um em cada ano do projeto, reunindo toda a equipe do projeto, incluindo os pesquisadores colaboradores do exterior. Além disso o projeto realizará seis jornadas de trabalho reunindo todos os pesquisadores da equipe atuando no Brasil.

8 Descrição da Infra-Estrutura Disponível

O IME-USP e o IEL-UNICAMP possuem uma infra-estrutura invejável.

Particularmente a biblioteca Carlos Benjamim de Lyra do IME-USP tornou-se referência nacional o que pode ser constatado em www.ime.usp.bib .

As instalações computacionais são atuais e atendem às necessidades do projeto com uma rede bastante desenvolvida e com suporte técnico/administrativo adequado (ver http://webinfo.ime.usp.br).

O grupo de pesquisadores, além da infra estrutura de uso comum das instituições pode contar ainda com o NUMEC - Núcleo de Modelagem Estocástica e Complexidade, instalação de 1800 m^2 dispostos em três pavimentos e um laboratório de informática com três computadores (sala de aproximadamente 8 m^2 no bloco A), ambos situados no IME-USP. Ainda, no IEL, há o Projeto Padrões Rítmicos, Fixação de Parâmetros e Mudança Lingüística que ocupa $50m^2$, distribuídos em três salas.

O equipamento, em rede, do projeto inclui:

- uma estação de trabalho DELL Poweredge 600C, Pentium IV, 2,4GHz, HD 40GB, monitor de 17 polegadas, rodando Linux, funcionando como servidora;
- 1 computadores DELL Optiplex GX100, Celeron 500MHz, monitor de 17 polegadas, rodando linux;
- 1 computador Mac Intosh G3;
- 2 computadores Pentium 1, rodando windows, com monitores de 17 polegadas;
- 2 impressoras a jato de tinta;
- 2 estabilizadores de voltagem.

Os pesquisadores podem contar com todos os recursos de reprografia e áudio-visual do IME-USP.

9 Projeção de benefícios complementares

Para a execução do projeto, projetamos as seguintes necessidades anuais de benefícios complementares:

- 15 passagens aéreas para os pesquisadores brasileiros realizarem visitas de trabalho aos centros onde trabalham os pesquisadores estrangeiros colaboradores do projeto;
- 225 diárias internacionais para os pesquisadores brasileiros realizarem visitas de trabalho aos centros onde trabalham os pesquisadores estrangeiros colaboradores do projeto;
- 12 passagens aéreas para os pesquisadores colaboradores estrangeiros realizarem visitas de pesquisa aos centros brasileiros;
- 180 diárias para os pesquisadores estrangeiros colaboradores do projeto participarem de atividades de pesquisa no Brasil;
- 6 bolsas de Iniciação científica ou Treinamento técnico em nível de graduação;
- 3 bolsas de Treinamento técnico em nível de pós-graduação;
- 3 bolsas de mestrado;
- 3 bolsas de doutorado;
- 3 bolsas de pós-doutorado.

10 Cadastros, Súmulas Curriculares e Curricula Lattes

11 Anexo: Artigos de Interesse

11.1 Artigo: Collet, P. ; Galves, A. e Lopes, A. (1995)

11.2 Artigo: Cassandro, M.; Collet, P. ; Galves, A. e Galves, C. (1999)

11.3 Artigo: Fernández, R. e Galves, A. (2000)

Referências

[1] Programa piccolo. *http://www.ime.usp.br/ tycho/prosody/piccolo/CD/program*, 2002.

[2] D. Abercrombie. *Elements of general phonetics.* Aldine, Chicago, 1967.

[3] M. Aizenman, J. Bricmont, and J. L. Lebowitz. Percolation of minoriy spins in high-dimensional Ising models. *J. Statist. Phys.*

[4] K. Athreya and P. Ney. A new approach to the limit theory of recurrent Markov chains. *Trans. Am. Math. Soc.*

[5] A. J. Baddeley, W. S. Kendall, and M. N. M. Van Lieshout. Quermass-interaction porcesses, 1996. Preprint.

[6] A. J. Baddeley and M. N. M. van Lieshout. Area-interaction point processes. *Ann. Inst. Statist. Math.*, 47(4):601–619, 1995.

[7] M. Barnsley, S. Demko, J. Elton, and J. Gerinomo. Invariant measures for Markov processes arising from iterated function systems with place-dependent probabilities. *Ann. Inst. H. Poincaré, Prob. Statist.*

[8] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*

[9] H. Berbee. Chains with complete connections: Uniqueness and Markov representation. *Prob. Th. Rel. Fields.*

[10] E. Bertin, J.-M. Billiot, and R. Drouilhet. Existence of 'nearest-neighbour' spatial Gibbs models. *Adv. Appl. Probab.*

[11] D. Blackwell. The entropy of functions on finite-state Markov chains. In *Trans. First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 13–20, Prague. Czechoslovak Akad. Sci.

[12] Hay J. Bod, R. and S. Jannedy. *Probabilistic Linguistics.* Mit Press, 2003.

[13] E. Borel. Sur les probabliités dénombrables et leurs applications arithmetiques. *Rend. Circ. Mat. Palermo.*

[14] C. Borgs and J. Z. Imbrie. A unified approach to phase diagrams in field theory and statistical mechanics. *Commun. in Math. Phys.*

[15] M. Bramson and S. A. Kalikow. Nonuniqueness in g-functions. *Israeli J. Math.*

[16] X. Bressaud, R. Fernández, and A. Galves. Speed of $\overline{d}$-convergence for Markov approximations of chains with complete connections. a coupling approach. *Stoch. Proc. and Appl.*

[17] X. Bressaud, R. Fernandez, and A. Galves. Decay of correlations for non holderian dynamics. a coupling approach. *Electron. J. Probab.*, 4, 1999.

[18] X. Bressaud, R. Fernández, and A. Galves. Decay of correlations for non Hölderian dynamics. a coupling approach. *Elect. J. Prob.*, 4, 1999b. (`http://www.math.washington.edu/~ejpecp/`).

[19] E. Brockmeyer, H. L. Halstrøm, and A. Jensen. The life and works of A. K. Erlang. *Trans. Danish Acad. Tech. Sci.*, 1948(2):277, 1948.

[20] E. Brockmeyer, H. L. Halstrøm, and A Jensen. The life and works of A. K. Erlang. *Acta Polytech. Scandinav. No.*, 287:277, 1960.

[21] D. C. Brydges. A short course on cluster expansions. In *Phénomènes critiques, systèmes aléatoires, théories de jauge, Part I, II (Les Houches, 1984)*, pages 129–183. North-Holland, Amsterdam-New York, 1986.

[22] P. Bühlmann and A. J. Wyner. Variable length Markov chains. *Ann. Statist.*

[23] H. Cai. A note on an exact sampling algorithm and Metropolis Markov chains, 1997. Preprint.

[24] M. Cassandro, P. Collet, A. Galves, and C. Galves. A statistical-physics approach to language acquisition and language change. *Physica A*, 263:427–437, 1999.

[25] H. Cohn. Limit theorems for systems with complete connections. *St. Cerc. Mat.*

[26] P. Collet, A. Galves, and A. Lopes. Maximum likelihood and minimum entropy identification of grammars. *Random and Computational Dynamics*, 3:241–256, 1995.

[27] P. Collet, A. Galves, and Schmitt. Repetition times for gibbsian sources. *Nonlinearity*, 12:1225–1237, 1999.

[28] F. Comets, R. Fernández, and P. A. Ferrari. Processes with long memory: Regenerative construction and perfect simulation. Preprint.

[29] F. Comets, R. Fernández, and P. A. Ferrari. Processes with long memory: Regenerative construction and perfect simulation. Preprint, can be retrieved from `http://xxx.lanl.gov/abs/math.PR/0009204`.

[30] D. Dacunha-Castelle, M. Duflo, and V. Genon-Catalot. *Exercices de Probabilités et Statistiques: 2-Problèmes à Temps Mobile.* Masson, Paris.

[31] R. L. Dobrushin. Central limit theorem for nonstationary markov chains. *Theor. Prob. and its Appl.*, 1:65–80 (Part I) and 329–383 (Part II), 1956.

[32] R. L. Dobrushin. Perturbation methods of the theory of Gibbsian fields. In *Lectures on probability theory and statistics (Saint-Flour, 1994)*, pages 1–66. Springer, Berlin, 1996.

[33] W. Doeblin. Exposé sur la théorie des chaînes simples constantes de Markoff à un nombre fini d'états. *Rev. Math. Union Interbalkanique*, pages 77–105.

[34] W. Doeblin. Remarques sur la théorie métrique des fractions continues. *Composition Math.*, 7:353–371.

[35] W. Doeblin. Sur deux problémes de M. Kolmogoroff concernant les chaînes dénombrables. *Bull. Soc. Math. Fr.*

[36] W. Doeblin and R. Fortet. Sur les chaînes à liaisons complétes. *Bull. Soc. Math. France.*

[37] D. Duarte, A. Galves, N.L. Garcia, and R. Maronna. The statistical analysis of acoustic correlates of speech rhythm. http://www.uni-bielefeld.de/complexity/duarte.pdf, 2001.

[38] R. Durrett. Ten lectures on particle systems. In *Lectures on probability theory (Saint-Flour, 1993)*, pages 97–201. Springer, Berlin, 1995.

[39] J. Elton and M. Piccioni. Iterated function systems arising from recursive estimation problems. *Prob. Th. Rel. Fields.*

[40] R. Fernández. Measures for lattice systems.

[41] R. Fernández. Random fields in lattices. the gibbsianness issue. *Resenhas do IME-USP*, 3:391–421.

[42] R. Fernández, P. A. Ferrari, and N. L. Garcia. Measures on contour, polymer or animal models. A probabilistic approach. *Markov Process. Related Fields*, 4(4):479–497, 1998. I Brazilian School in Probability (Rio de Janeiro, 1997).

[43] R. Fernández, P. A. Ferrari, and N. L. Garcia. Loss network representation of Peierls contours, 1999. Preprint.

[44] R. Fernández, P. A. Ferrari, and N. L. Garcia. Perfect simulation for interacting point processes, loss networks and ising models, 1999. Preprint.

[45] R. Fernández and A. Galves. Identifying features in the presence of competing evidence. the case of first-language acquisition. *World Sci. Ser. Appl. Anal.*, pages 52–62, 2000.

[46] R. Fernández and A. Galves. Markov approximations of chains of infinite order. *Bol. Soc. Brasil. Mat.*, 33:to appear, 2002.

[47] R. Fernández, P. Ferrari, and A. Galves. Coupling, renewal and perfect simulations of chain of infinite order. In *5a Escola Brasileira de Probabilidades*, 2001. http://www.ime.usp.br/ galves/livro/revised2.pdf.

[48] R. Fernández and A. Galves. Identifying features in the presence of competing evidence. the case of first-language acquisition. *World Sci. Ser. Appl. Anal.*, pages 52–62, 2000.

[49] P. A. Ferrari. Ergodicity for spin systems with stirrings. *Ann. Probab.*

[50] P. A. Ferrari and A. Galves. *Acoplamentos e Processos Estocásticos.* IMPA, Rio de Janeiro, Brazil.

[51] P. A. Ferrari and A. Galves. *Construction of Stochastic processes, Coupling and Regeneration.* Facultad de Ciencias de la Universidad de los Andes, Mérida, Venezuela.

[52] P. A. Ferrari and N. L. Garcia. One-dimensional loss networks and conditioned $M/G/\infty$ queues. *J. Appl. Probab.*, 35(4):963–975, 1998.

[53] P. A. Ferrari, A. Maass, S. Martíinez, and P. Ney. Cesàro mean distribution of group automata starting from measures with summable decay. To be published in *Ergodic Th. Dyn. Syst.*

[54] P. A. Ferrari, A. Maass, S. Martínez, and P. Ney. Cèsaro mean distribution of group automata starting from measures with summable decay. To be published in *Ergodic Th. Dyn. Syst.*

[55] J. A. Fill. An interruptible algorithm for perfect sampling via Markov chains. *Ann. Appl. Probab.*, 8(1):131–162, 1998.

[56] J. A. Fill. The move-to-front rule: a case study for two perfect sampling algorithms. *Probab. Engrg. Inform. Sci.*, 12(3):283–302, 1998.

[57] J. A. Fill, M. Machida, D. J. Murdoch, and J. S. Rosenthal. Extension of Fill's perfect rejection sampling algorithm to general chains, 1999. Preprint.

[58] S. G. Foss and R. L. Tweedie. Perfect simulation and backward coupling. *Comm. Statist. Stochastic Models*, 14(1-2):187–203, 1998. Special issue in honor of Marcel F. Neuts.

[59] S. G. Foss and R. L. Tweedie. Perfect simulation and backward coupling. *Stochastic Models*, 14(1-2):187–203, 1998.

[60] S. G. Foss, R. L. Tweedie, and J. N. Corcoran. Simulating the invariant measures of Markov chains using backward coupling at regeneration times. *Probab. Engrg. Inform. Sci.*, 12(3):303–320, 1998.

[61] S. G. Foss, R. L. Tweedie, and J. N. Corcoran. Simulating the invariant measures of Markov chains using backward coupling at regeneration times. *Probability in the Engineering and Informational Sciences*, 12:303–320, 1998.

[62] N. A. Friedman and D. S. Ornstein. On isomorphism of weak bernoulli transformations. *Advances in Math.*, 5:365–394.

[63] A. Galves, J. Garcia, D. Duarte, and C. Galves. Sonority as a basis for rhythmic class discrimination. In *Speech Prosody 2002*, Aix-en-Provence, 2002. http://www.lpl.univ-aix.fr/sp2002.

[64] H.-O. Georgii. Phase transitions and percolation in Gibbsian particle models. Preprint, can be retrieved from `http://dimacs.rutgers.edu/~dbwilson/exact`.

[65] H.-O. Georgii. *Gibbs Measures and Phase Transitions*. Walter de Gruyter (de Gruyter Studies in Mathematics, Vol. 9), Berlin–New York, 1988.

[66] H.-O. Georgii and Häggstr̈om. Phase transition in continuum Potts models. *Commun. Math. Phys.*

[67] S. Grigorescu and Gh. Popescu. A central limit theorem for a class of Markov chains. In *Proceedings of the 4th Conference on Probability Theory*, Bucarest. Academiei Bucaresti.

[68] X. Guyon. *Random Fields on a Network*. Springer-Verlag (Probabilities and its Applications), New York.

[69] O. Häggström and K. Nelander. On exact simulation of Markov random fields using coupling from the past. *Scandinavian Journal of Statistics*, 1997. To appear.

[70] O. Häggström and K. Nelander. Exact sampling from anti-monotone systems. *Statistica Neerlandica*, 52:360–380, 1998.

[71] O. Häggström and J. E. Steif. Propp-Wilson algorithms and finitary codings for high noise Markov random fields, 1998. Preprint.

[72] O. Häggström, M. N. M. van Lieshout, and J. Møller. Characterisation results and Markov chain Monte Carlo algorithms including exact simulation for some spatial point processes. Technical Report R-96-2040, Aalborg University, 1996. To appear in *Bernoulli*.

[73] T. Harris. Nearest-neighbor Markov interaction processes on multidimensional lattices. *Advanves in Math.*, 9:66–89.

[74] T. E. Harris. Additive set-valued markov processes and graphical methods. *Ann. Probability*, 6:355–378.

[75] T. E. Harris. The existence of stationary measures for certain Markov processes. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistical and Probability*, pages 113–124, Berkeley. University of California Press.

[76] T. E. Harris. On chains of infinite order. *Pacific J. Math.*, 5:707–24.

[77] T. E. Harris. Nearest-neighbor Markov interaction processes on multidimensional lattices. *Advances in Math.*, 9:66–89, 1972.

[78] John A. Hawkins and Murray Gell-Mann, editors. *The evolution of human languages*. SFI Studies in the Sciences of Complexity. Addison Wexley Longman, 1992.

[79] U. Herkenrath and R. Theodorescu. General control systems. *Information Sci.*

[80] M. Huber. Efficient exact sampling from the Ising model using Swendsen-Wang, 1998. A two-page version appeared in *Tenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Preprint.

[81] M. Iosifescu. A coupling method in the theory of dependence with complete connections according to Doeblin. *Rev. Roum. Math. Pures et Appl.*

[82] M. Iosifescu. On the asymptotic behaviour of chains with complete connections. *Comunicările Acad. RPR.*

[83] M. Iosifescu. Recent advances in the metric theory of continued fractions. In *Trans. 8th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (Prague, 1978)*, volume A, pages 27–40, Dordrecht-Boston. Reidel.

[84] M. Iosifescu and S. Grigorescu. *Dependence with Complete Connections and its Applications.* Cambridge University Press, Cambridge, UK.

[85] M. Iosifescu and R. Theodorescu. *Random Processes and Learning.* Springer-Verlag, Berlin.

[86] F. Jelinek. *Statistical Methods of Speech Recognition.* MIT university Press, Boston.

[87] M. Jerrum and A. Sinclair. The Markov chain Monte Carlo method: an approach to approximate counting and integration. In D. S. Hochbaum, editor, *Approximation Algorithms for NP-hard Problems*, pages 482–520, Boston. PWS Publishing.

[88] T. Kaijser. Another central limit theorem for random systems with complete connections. *Rev. Roumaine Math. Pure Appl.*

[89] T. Kaijser. A limit theorem for partially observed Markov chains. *Ann. Prob.*, 3:677–96.

[90] T. Kaijser. On a new contraction condition for random systems with complete connections. *Rev. Roum. Math. Pures et Appl.*

[91] T. Kaijser. On a theorem of Karlin. *Acta Applicandae Mathematicae.*

[92] M. Keane. Strongly mixing g-measures. *Inventiones Math.*

[93] M. Keane. Sur les mesures invariants d'un recouvrement régulier. *C. R. Acad. Sci. Paris.*

[94] W. S. Kendall. On some weighted Boolean models. In D. Jeulin, editor, *Proceedings of the International Symposium on Advances in Theory and Applications of Random Sets (Fontainebleau, 1996)*, pages 105–120. World Sci. Publishing, River Edge, NJ, 1997.

[95] W. S. Kendall. Perfect simulation for spatial point processes. In *Bulletin of the International Statistical Institute* 51^{st} *Session, Istanbul (August 1997)*, volume 3, pages 163–166, 1997.

[96] W. S. Kendall. Perfect simulation for the area-interaction point process. In L. Accardi and C. C. Heyde, editors, *Probability Towards 2000*, pages 218–234. Springer, 1998.

[97] W. S. Kendall and J. Møller. Perfect Metropolis-Hastings simulation of locally stable point processes, 1999. Preprint.

[98] C. Kipnis and C. Landim. *Scaling Limits of Interacting Particle Systems.* Springer-Verlag, Heidelberg, etc.

[99] A. N. Kolmogorov and N. G. Rychkova. Russian poetry rhythm analysis and probability theory. *Theory of Probability and its applications*, 44:375–385, 2000.

[100] R. Kotecký and D. Preiss. Cluster expansion for abstract polymer models. *Comm. Math. Phys.*, 103(3):491–498, 1986.

[101] O. K. Kozlov. Gibbs description of a system of random variables. *Probl. Inform. Transmission*, 10:258–265, 1974.

[102] S. P. Lalley. Regeneration in one-dimensional Gibbs states and chains with complete connections. *Resenhas IME-USP*, 4:249–80.

[103] S. P. Lalley. Regeneration representation for one-dimensional Gibbs states. *Ann. Prob.*

[104] J. L. Lebowitz and G. Gallavotti. Phase transitions in binary lattice gases. *J. Math. Phys.*

[105] J. L. Lebowitz and A. E. Mazel. Improved Peierls argument for high-dimensional Ising models. *J. Statist. Phys.*

[106] F. Ledrappier. Principe variationnel et systèmes dynamiques symboliques. *Z. Wahrscheinlichkeitstheorie verw. Gebiete.*

[107] T. M. Liggett. The coupling technique in interacting particle systems. Doeblin and modern probability. *Contemp. Math.*

[108] T. M. Liggett. *Interacting Particle Systems.* Springer-Verlag, Berlin.

[109] T. M. Liggett. *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes.* Springer-Verlag, Berlin.

[110] T. Lindvall. *Lectures on the Coupling Method.* Wiley, New York.

[111] T. Lindvall. W. Doeblin 1915–1940. *Ann. Prob.*

[112] J. Lloyd. *Speech signal in telephony.* unknown, London, 1940.

[113] C. Maes and S. B. Shlosman. Ergodicity of probabilistic cellular automata: A constructive criterion. *Commun. Math. Phys.*

[114] C. Maes and S. B. Shlosman. When is an interacting particle system ergodic? *Commun. Math. Phys.*

[115] K. Marton. Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration. Ann. Probab.

[116] P. McCullagh and J. A. Nelder. *Generalized linear Models* (2nd Edition). Chapman-Hall, London.

[117] J. Mehler, E. Dupoux, T. Nazzi, and Dehane-Lambertz. Coping with linguistic diversity: the infant's viewpoint. In J.L Morgan and K.D Demuth, editors, *Signal to sintax: bootstrapping from speech to grammar in early acquisition.* verificar, 1996.

[118] Gh. Mihoc. The limit law for sums of vector-valued random-variables forming a multiple stationary chain with complete connections. *Comunicãrile Acad. RPR.*

[119] J. Møller. On the rate of convergence of spatial birth-and-death processes. *Ann. Inst. Statist. Math.*, 41(3):565–581, 1989.

[120] J. Møller. Markov chain Monte Carlo and spatial point processes. In W. S. Kendall, O. E. Barndorff-Nielsen, and M. N. M. van Lieshout, editors, *Stochastic Geometry: Likelihood and Computation*, Monographs on Statistics and Applied Probability #80, pages 141–172. Chapman and Hall / CRC Press, 1998.

[121] J. Møller. Perfect simulation of conditionally specified models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(1):251–264, 1999.

[122] J. Møller and K. Schladitz. Extensions of Fill's algorithm for perfect simulation. *Journal of the Royal Statistical Society* B, 61, 1998. To appear.

[123] D. J. Murdoch and P. J. Green. Exact sampling from a continuous state space. *Scand. J. Statist.*, 25(3):483–502, 1998.

[124] D. J. Murdoch and P. J. Green. Exact sampling from a continuous state space. *Scandinavian Journal of Statistics*, 25(3):483–502, 1998.

[125] F. Norman. An ergodic theorem for evolution in a random environment. *J. Appl. Prob.*

[126] F. Norman. *Markov Processes and Learning Models.* Academic Press, New York.

[127] F. Norman. Markovian learning processes. *SIAM Review.*

[128] E. Nummelin. A splitting technique for Harris recurrent Markov chains. *Z. Wahrscheinlichkeitstheorie verw. Gebiete.*

[129] E. Nummelin and P. Ney. Regeneration for chains with infinite memory. *Prob. Th. Rel. Fields.*

[130] O. Onicescu and G. Mihoc. Le comportement asymptotique des chaînes à liaisons complètes. *Disq. Math. Phys.*, 1:61–2.

[131] O. Onicescu and G. Mihoc. Sur les chaînes de variables statistiques. *Bull. Sci., Math.*

[132] O. Onicescu and G. Mihoc. Sur les chaînes statistiques. *C. R. Acad. Sci. Paris.*

[133] D. S. Ornstein. *Ergodic Theory, Randomness and Dynamical Systems.* Yale University Press (Yale Mathematical Monographs 5).

[134] D. S. Ornstein and B. Weiss. How sampling reveals a process. *Ann. Prob.*

[135] W. Parry and M. Pollicott. Zeta functions and the periodic structure of hyperbolic dynamics. *Asterisque*, 187-188.

[136] K. L. Pike. *The intonation of American English.* University of Michigan Press, Ann Arbor, 1945.

[137] J. Propp. Generating random elements of a finite distributive lattice. *Electronic Journal of Combinatorics*, 4(2), 1997. Paper #R15.

[138] J. Propp and D. Wilson. Coupling from the past: a user's guide. In D. Aldous and J. Propp, editors, *Microsurveys in Discrete Probability*, volume 41 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 181–192. American Mathematical Society, 1998.

[139] J. G. Propp and D. Wilson. Coupling from the past: a user's guide. In *Microsurveys in discrete probability (Princeton, NJ, 1997)*, pages 181–192. Amer. Math. Soc., Providence, RI, 1998.

[140] J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*, volume 9, pages 223–252, 1996.

[141] J. G. Propp and D. B. Wilson. How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *J. Algorithms*, 27(2):170–217, 1998. 7th Annual ACM-SIAM Symposium on Discrete Algorithms (Atlanta, GA, 1996).

[142] H. Pruscha and R. Theodorescu. Functions of event variables of a random system with complete connections. *J. Multivariate Anal.*, 7:336–62.

[143] P. Ferrari R. Fernández and A. Galves. Coupling, renewal and perfect simulations of chain of infinite order. 5^a *Escola Brasileira de Probabilidades*, 2001.

[144] A. Raftery. A model for high-order Markov chains. *J. R. Statist. Soc. B.*

[145] A. Raftery. A new model for discrete-valued time series: autocorrelations and extensions. *Rass. Met. Statist. Appl.*, 3–4:149–162.

[146] A. Raftery and A. Tavaré. Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model. *Appl. Statist.*

[147] F. Ramus, M. Nespor, and J. Mehler. Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3):265–292, 1999.

[148] D. J. Rudolph and G. Schwarz. The limit in $\overline{d}$ of multi-step Markov chains. Israel Journal of Math.

[149] D. Ruelle. Existence of a phase transition in a continuous classical system. *Phys. Rev. Lett.*

[150] F. Sandalo, A. Mandel, M.B. Abaurre, and C. Galves. The sotaq optimality based computer program and secondary stress in two varieties of portuguese. *Probus*, no prelo. http://www.ime.usp.br/ tycho/prosody/sotaq.

[151] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal.*

[152] P. Shields. Cutting and stacking: A method for constructing stationary processes. *IEEE Trans. Inform. Theory*, IT-37:1605–17.

[153] F. Spitzer. Interaction of Markov processes. *Advanves in Math.*, 5:246–290.

[154] David J. Strauss. A model for clustering. *Biometrika*, 62(2):467–475, 1975.

[155] E. Thönnes. Perfect simulation of some point processes for the impatient user. *Advances in Applied Probability, Stochastic Geometry and Statistical Applications*, 1997. To appear.

[156] H. Thorrison. *Coupling, Stationarity and Regeneration.* Springer-Verlag, Heidelberg.

[157] A. C. D. van Enter, R. Fernández, and A. D. Sokal. Regularity properties and pathologies of position-space renormalization-group transformations: scope and limitations of Gibbsian theory. *J. Stat. Phys.*, 72:879–1167, 1993.

[158] P. Walters. Ruelle's operator theorem and g-measures. *Trans. Amer. Math. Soc.*

[159] B. Widow and J. S. Rowlinson. New model for the study of liquid-vapor phase transitions. *J. Chem. Phys.*

[160] D. B. Wilson. Annotated bibliography of perfectly random sampling with Markov chains. In D. Aldous and J. Propp, editors, *Microsurveys in Discrete Probability*, volume 41 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 209–220. American Mathematical Society, 1998. Updated versions can be found at `http://dimacs.rutgers.edu/~dbwilson/exact`.

[161] R. Fernandez X. Bressaud and A. Galves. Speed of $\overline{d}$-convergence for markov approximations of chains with complete connections. a coupling approach. *Stochastic Process. Appl.*, no. 1:127–138, 1999.