Markovian modeling of the stress contours of Brazilian and European Portuguese

C. Dorea[1] , A. Galves[2] , E. Kira[2] and A. Pereira Alencar[2]

Abstract. This work addresses the question of modeling the stress contours of Brazilian and Modern European Portuguese as high order Markov chains. We discuss three criteria to select the order of the chain: the *Akaike's Information Criterion*, the *Bayesian Information Criterion* and the *Minimum Entropy Criterion*. A statistical analysis of a sample of spontaneous speech from both dialects indicates that the corresponding Markov chains are of different order.

1 Introduction

The human brain codes the prosodic features which are present in human speech as a sequence of elements belonging to a finite set and evolving in time as a stochastic process which is both stationary and ergodic. *Stationary* means that the process is homogeneous in time. *Ergodic* means that these features do not depend on the particular sample we are considering.

The main goal of the present work is to identify the stochastic processes which are responsible for the stress contours of European Portuguese (EP) and Brazilian Portuguese (BP). This identification should be able to put in evidence what these processes have in common and in what they differ. Also a critical review of the available methods to identify the order of a chain is presented.

The data set under analysis is part of a corpus organized by M. B. Abaurre and co-workers [1] at Campinas State University. This corpus is constituted of phonetic transcriptions of sentences produced under various circumstances by speakers of the two dialects of Portuguese under consideration. The stress contours of these sentences were codified by human means. By stress contour we mean an ordered sequence of stressed and non stressed elements between two phrase boundaries. We focus on the sequence of distances between consecutive stressed elements. It seems reasonable to conjecture that the sequence of these distances between two boundaries symbols behave as a Markov chain of high order. Eventually this order could even be zero, which is the independent case.

Besides the estimation of the transition probabilities, we discuss three different criteria to estimate the order of a chain, namely the Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Minimum Entropy Criterion.

Strange as it may appears, there are few papers ([6], [9], [10]) dealing with the estimation of the order of the chain. Besides the classical papers by Anderson and Goodman ([2]) and Billingsley ([3]), most of the papers are related to probabilistic modeling of DNA sequences

[1]Departamento de Matemática, Universidade de Brasília, 70 919-970 Brasília DF, cdorea@mat.unb.br
[2]Departamento de Estatística, Universidade de São Paulo, Caixa Postal 66281, 05315-970 São Paulo, SP; galves@ime.usp.br, betikira@ime.usp.br, lane@ime.usp.br, respectively

(cf. [9] and [12]). In computational linguistics, the idea of modeling the sequence of prosodic phrases as a Markov chain is sketched in [11] but without discussing the issues addressed here. In particular, as far as we know, there are no comparative studies between EP and BP using Markovian models.

This paper is organized as follows. The statistical methods used are described in Section 2 and the analysis of the data is carried out in Section 3.

2 Statistical Methods

Let $(Y_n)_{n\geq 0}$ a Markov chain of order k on the finite set $S = \{1, 2, \dots, m\}$. If the chain is stationary, it is characterized by its transition probability matrix having entries of the form

$$p(y_k \mid y_0, y_1, \dots, y_{k-1}) = P(Y_{t+k} = y_k \mid Y_t = y_0, Y_{t+1} = y_1, \dots, Y_{t+k-1} = y_{k-1}) \quad t = 1, 2, \dots \tag{1}$$

where $y_j \in S$, $j = 0, 1, \dots, k$.

In our case the sequence of distances between stressed elements that are not separated by a boundary is to be model by a high order Markov chain. Since in the data set the maximum distance between two stressed elements was found to be three and only very few observations with distance four. Therefore for our discussion m was taken to be three.

In the sequence we present log–likelihood estimators of the transition probabilities and the criteria considered for the estimation of the order of the chain.

2.1 The log–likelihood estimators

Let us assume (as in [3]) that our sample is a string of observations $(a_0, a_1, \dots, a_n)$ of the Markov chain $(Y_n)_{n\geq 0}$ of order k with state space $S = \{1, \dots, m\}$.

We can associate with the process (Y_n) a derived chain (Y'_n) with state space $\mathcal{S}' = \{(z_1, \dots, z_k) : z_j \in S, \ j = 1, \dots, k\}$. The chain (Y'_n) is said to be in state $v \in \mathcal{S}'$ at time t if $v = (y_t, y_{t+1}, \dots, y_{t+k-1})$, where y_u is the state of the original chain (Y_n) at time u. It follows that the derived chain is a first-order Markov chain with transition matrix $P = (p_{vw})$ where for $v = (v_1, \dots, v_k)$ and $w = (w_1, \dots, w_k)$ we have,

$$p_{vw} = \begin{cases} p(w_k \mid v_1, \dots, v_k) & \text{if } w_i = v_{i+1}, \quad i = 1, \dots, k-1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus inference can be done on the chain (Y'_n) instead of the original chain (Y_n).

To estimate the transition probabilities we may write $\mathcal{S}' = \{1, 2, \dots, s\}$ with $s = m^k$. For a given realization of the process (Y'_n) up to time n, say, $(a'_0, a'_1, \dots, a'_n)$ let f_{ij} denote the number of integers ℓ such that $0 \leq \ell \leq n$, $a'_\ell = i$ and $a'_{\ell+1} = j$, with $i, j \in \mathcal{S}'$. That is, f_{ij} represents the transition counts from state i to j in the realization $(a'_0, a'_1, \dots, a'_n)$. Then the log-likelihood becomes

$$\log L' = \sum_D f_{ij} \log p_{ij} \quad \text{where} \quad D = \{(i,j) : p_{ij} > 0\}.$$

To find the maximum likelihood estimator $\hat{p}_{vw}$ one need to maximize $\log L'$ subject to the constraint $\sum_w p_{vw} = 1$. For stationary and ergodic chains this can be accomplished using the Lagrange multipliers:

$$\hat{p}_{ij} = \frac{f_{ij}}{f_i} \quad \text{with} \quad f_i = \sum_{j \in \mathcal{S}'} f_{ij}. \tag{2}$$

Moreover, it can be shown that as $n \to \infty$,

$$n^{1/2} \left(\frac{f_{ij}}{f_i} - p_{ij} \right) \longrightarrow 0 \quad \text{(in probability)}.$$

Now let $\xi_{ij} = (f_{ij} - f_i p_{ij})/f_i^{1/2}$. Then the chi-square methods applicable to the multinomial case can be carried over to the Markov case. And it can be shown that if the chain is stationary and ergodic then the distribution of the s^2-dimension random vector (ξ_{ij}) converges as $n \to \infty$ to the normal distribution with zero mean and covariance matrix $E\xi_{ij}\xi_{r\ell} = \delta_{ir}(\delta_{j\ell}p_{ij} - p_{ij}p_{i\ell})$ where δ_{ir} is the Dirac function. It follows that for $n \to \infty$,

$$U_n(i) = \sum_j \frac{(f_{ij} - f_i\, p_{ij})^2}{f_i\, p_{ij}} \longrightarrow \chi^2\text{-distribution}$$

with $d_i - 1$ degrees of freedom ($d_i = card\{j : p_{ij} > 0\}$). Since for $i = 1, \ldots, s$ the statistics $U_n(i)$ are asymptotically independent we also have $U_n \to \chi^2$, with $d - s$ degrees of freedom, where

$$U_n = \sum_{i,j} \frac{(f_{ij} - f_i\, p_{ij})^2}{f_i\, p_{ij}}$$

and $d = \sum_i d_i$. The statistics U_n is useful for testing whether the transition probabilities of the chain have specified values p^o_{ij} (goodness of fit test).

The corresponding statistics for the process (Y_n) then becomes

$$\hat{p}(a_k \mid a_0 \ldots a_{k-1}) = \frac{f_{a_0 \ldots a_k}}{f_{a_0 \ldots a_{k-1}}} \ , \tag{3}$$

$$\hat{L_k} = \prod \hat{p}(a_k \mid a_0, a_1, \ldots, a_{k-1})^{f_{a_0, \ldots, a_k}} \ , \tag{4}$$

$$U_n \quad = \sum_{a_0 \ldots a_k} \frac{(f_{a_0 \ldots a_k} - f_{a_0 \ldots a_{k-1}}\ p(a_k \mid a_0 \ldots a_{k-1}))^2}{f_{a_0 \ldots a_{k-1}} p(a_k \mid a_0 \ldots a_{k-1})} \ , \tag{5}$$

where $f_{a_0 \ldots a_k}$ is the number of t, $1 \le t \le n-1$ such that $(y_t, \ldots y_{t+k}) = (a_0, \ldots, a_k)$ and $f_{a_0 \ldots a_{k-1}} = \sum_{a_k} f_{a_0 \ldots a_k}$.

Moreover,

$$2\log\frac{\hat{L}_{k+1}}{\hat{L}_k}\longrightarrow\chi^2 \tag{6}$$

with $m^k(m-1)^2$ degrees of freedom.

2.2 The AIC and the BIC

We recall that the set of high order Markov chains is dense in the set of stationary stochastic processes. Therefore, the fitness of a Markov chain model to the data can be improved by increasing its order k. However the number of independent parameters $\gamma=(m-1)m^k$ grows exponentially on k, so that one needs to take this fact into account by penalizing a large value of k when it is not needed. This is precisely the aim of the *Akaike's Information* and *Bayesian Information Criteria* (for further details see [9], [6], [10].)

For a fixed k, let the AIC(k) and BIC(k) coefficients be defined by

$$\mathrm{AIC}(k)=-2\log\hat{L_k}+2\gamma \text{ and} \tag{7}$$

$$\mathrm{BIC}(k)=-2\log\hat{L_k}+\gamma\log n\,. \tag{8}$$

where n is the size of the observed sequence and $\hat{L_k}$ is given by (4).

The order of the chain is taken to be the value $\hat{k}$ that minimizes (7) for AIC and (8) for BIC.

The BIC criterion was proposed by [10] and further studied by [6] as an alternative to AIC.

The BIC amounts to maximize the asymptotic posterior distribution of the parameters when the prior distribution is of uniform type, i.e. the set of transition probabilities is assumed to have an independent Dirichlet distribution with all parameters unity.

An heuristic interpretation of (8) which conciliates the Bayesian and the frequentist points of view is the following. Let us assume that the *a priori* distribution is a product measure of k uniform distributions on a finite subset obtained by considering a partition of the interval [0,1] in subintervals of length u, i.e. $\{[0,u),[u,2u),\ldots,[[\frac{1}{u}]u,1]\}$. Taking into account the Central Limit Theorem, if we have a sample of length n then the precision we can expect in the result is of order $\frac{1}{\sqrt{n}}$. So it is reasonable to take $u=1/\sqrt{n}$. Therefore, we can take the *a priori* distribution as

$$P_{priori}(p_1,\ldots,p_\gamma)=(\frac{1}{\sqrt{n}})^\gamma\,.$$

Using Bayes formula and taking the logarithm of the *a posteriori* distribution we obtain directly the expression (8).

2.3 The Minimum Entropy Criterion

Let $(a_0, a_1, \ldots, a_n)$ be a sample produced by a Markov chain $(Y_n)_{n\geq 0}$ be a Markov chain of order ℓ taking values in a finite set $S = \{1, 2, ..., m\}$. For a fixed k, let

$$\hat{h}_k = \hat{h}_k((a_0, a_1, \ldots, a_n)) = -\sum \hat{p}(a_0, a_1, \ldots, a_{k-1}, a_k) \log \hat{p}(a_k \mid a_0, a_1, \ldots, a_{k-1}) \ . \tag{9}$$

We remark that if $k = \ell$ then (9) is an estimate of the *entropy* of the chain (for more details on the notion of entropy see [4] and [7]).

The *Minimum Entropy Criterion* takes the order of the chain as the minimum value of k that minimizes h_k with a desirable level of significance. This criterion is justified by the following propositions.

Proposition 2.1 *Let $(Z_n)_{n\in\mathbf{Z}}$ be a stationary process taking values in a finite set S. Let us define h_0 as $-E[\log P(Z_0)]$, and for each $k \geq 1$,*

$$h_k = -E[\log P(Z_0 \mid Z_{-k}, ..., Z_{-1})] \ .$$

Then the sequence $(h_k)_k$ is monotone and decreasing. Moreover, $h_{\ell-1} > h_\ell = h_j$, for some ℓ and for all $j \geq \ell$, if and only if (Z_n) is a Markov chain of order ℓ.

Proof: Using the short-hand notation

$$P(Z_{k+1} = z_{k+1} \mid Z_0 = z_0, ..., Z_k = z_k) = p(z_{k+1} \mid z_0, ..., z_k)$$

and $P(Z_0 = z_0, ..., Z_k = z_k) = p(z_0, ..., z_k)$, we first rewrite h_{k+1} as

$$\begin{aligned} h_{k+1} &= -E[\log P(Z_{k+1} \mid Z_0, ..., Z_k)] \\ &= -\sum_{z_0,\ldots,z_{k+1}} p(z_0, ..., z_{k+1}) \log p(z_{k+1} \mid z_0, ..., z_k) \\ &= \sum_{z_1,\ldots,z_{k+1}} p(z_1, ..., z_{k+1}) \sum_{z_0} \frac{p(z_0, ..., z_{k+1})}{p(z_1, ..., z_{k+1})} \log \frac{p(z_0, ..., z_k)}{p(z_0, ..., z_{k+1})} \ . \end{aligned}$$

Using Jensen's inequality we obtain

$$\begin{aligned} h_{k+1} &\leq \sum_{z_1,\ldots,z_{k+1}} p(z_1, ..., z_{k+1}) \log \sum_{z_0} \frac{p(z_0, ..., z_k)}{p(z_1, ..., z_{k+1})} \\ &= -\sum_{z_1,\ldots,z_{k+1}} p(z_1, ..., z_{k+1}) \log p(z_{k+1} \mid z_1, ..., z_k) = h_k \ . \end{aligned}$$

Note that equality in the above expression holds if and only if

$$p(z_{k+1} \mid z_0, ..., z_k) = p(z_{k+1} \mid z_1, ..., z_k), \forall \ (z_0, ..., z_k, z_{k+1}) \ .$$

$\diamond$

Proposition 2.1 implies that if $(Y_n)_{n\geq 0}$ is a Markov chain, its order ℓ is given by

$$\ell = \min\{k \geq 0 : h_{k+1} = h_k\} \ .$$

Proposition 2.2 *If* $(Y_n)_{n\geq 0}$ *is a chain of order* k *then*

$$-2n(\hat{h}_{k+1} - \hat{h}_k) \longrightarrow \chi^2$$

with $m^k(m-1)^2$ *degrees of freedom, where* $\hat{h}_k$ *is defined by (9).*

Proof: The result follows immediately from (6) and the relation

$$-2\log\frac{\hat{L}_k}{\hat{L}_{k+1}} = 2n(\hat{h}_k - \hat{h}_{k+1}) \ ,$$

where n is the sample size. $\diamond$

3 Application to prosodic data of Brazilian Portuguese and European Portuguese

3.1 Description of the data set

In order to compare the Brazilian Portuguese and the European Portuguese, we analyzed data sets extracted from the data bank organized by M. Bernadete Abaurre and co-workers, at IEL-UNICAMP, part of which was presented in [1].

This data bank was prepared using transcriptions of cassette tapes of samples of spontaneous speech by Brazilian and European Portuguese native speakers. A sample of European Portuguese was extracted from the data bank *Português Fundamental* (cf. [8]). The phonetic transcriptions for both dialects of Portuguese and their codification were prepared by Abaurre and co-workers.

The data was codified with symbols B, 1, and 0, denoting the boundaries of intonational phrases, stressed and non-stressed elements of the sentences, respectively. The statistical analysis focused on the sequence of distances between consecutive stressed elements. Distances between stressed elements separated by a boundary B, were disconsidered, in order to avoid a possible bias introduced by the presence of the boundaries.

For illustration we give a short example of spoken EP and the corresponding phonetic transcription: "*... e então achei que devia ter lençóis de banho ou toalhas todas azuis ou todas amarelas. E ela: a falta de sentido prático que um homem tem. ...*".

> ... (B) e(1) en(0)tão(1) a(0)chei(1) que(1) devia(1) ter(1) len(0)çóis(1) de(0) ba(1)nho ou(0) toa(1)lhas(0) to(1)das(0) a(0)zuis(1) ou(1) to(1)das(0) a(1)ma(0)re(1)las(0). (B) e(0) e(1)la a(0) fal(1)ta de(0) sen(0)ti(1)do(0) práti(1)co que um(0) ho(1)mem(0) tem(1).(B) ...

Which give us the following transcription:

B 1 0 1 0 1 1 1 1 0 1 0 1 0 1 0 1 0 0 1 1 1 0 1 0 1 0 B 0 1 0 1 0 0 1 0 1 0 1 0 1 B

For the first stress contour: B 1 0 1 ... 1 0 B the corresponding sequence of distances between consecutive stressed elements is

$$(2,2,1,1,1,2,2,2,2,3,1,1,2,2).$$

For the remaining string B 0 1 0 ... 0 1 B we have (2, 3, 2, 2, 2).

Since we will be testing the hypothesis that the sequences were produced by Markov chains up to order 2, it is necessary to consider stress contours with at least four successive stressed elements. Therefore intonational phrases with less than four stressed elements were eliminated in our analysis. And this procedure reduced considerably the size of the original sample.

The criteria were applied to a sample of Brazilian Portuguese of final size 76 and a sample of European Portuguese of final size 219.

3.2 Results and discussion

Using the expression given in (3) the estimates of the transition probabilities can be computed. For the independent case ($k = 0$) these estimates are displayed at Tables 1 and 2. The corresponding estimates for $k = 1$ and $k = 2$ are shown at Tables 3 and 4, and Tables 5 and 6, respectively.

Table 1: Probability distribution for BP ($k = 0$).

distance	1	2	3
probability	0.132	0.658	0.211

Table 2: Probability distribution for EP ($k = 0$).

distance	1	2	3
probability	0.146	0.731	0.123

Table 3: Transition probabilities for BP ($k = 1$).

$$\begin{pmatrix} 0.125 & 0.750 & 0.125 \\ 0.118 & 0.667 & 0.216 \\ 0.176 & 0.588 & 0.235 \end{pmatrix}$$

Table 4: Transition probabilities for EP ($k = 1$).

$$\begin{pmatrix} 0.289 & 0.658 & 0.053 \\ 0.112 & 0.737 & 0.151 \\ 0.138 & 0.793 & 0.069 \end{pmatrix}$$

For each i and j, $i, j = 1, 2, 3$ the entry (i, j) of the matrices above represents the transition probability $p_{ij} = p(j \mid i)$.

Table 5: Transition probabilities for BP ($k = 2$).

l / (i,j)	1	2	3
11	0.000	0.000	0.000
12	0.167	0.500	0.333
13	0.000	0.667	0.333
21	0.143	0.714	0.143
22	0.103	0.621	0.276
23	0.300	0.600	0.100
31	0.000	1.000	0.000
32	0.125	0.813	0.063
33	0.000	0.500	0.500

Table 6: Transition probabilities for EP ($k = 2$).

l / (i,j)	1	2	3
11	0.333	0.583	0.083
12	0.148	0.815	0.037
13	0.000	1.000	0.000
21	0.238	0.762	0.000
22	0.104	0.736	0.160
23	0.160	0.760	0.080
31	0.400	0.400	0.200
32	0.105	0.632	0.263
33	0.000	1.000	0.000

In the above matrices, the entry corresponding to row (i, j) and column l indicates the transition probability $p(l \mid i, j)$.

Using (7), (8), and (9) the AIC, the BIC, and the entropies differences with the respective p-values were calculated for the BP and EP data sets and are presented on Tables 7 and 8. The last two columns of these tables contains the p-value and the degrees of freedom (d.f.) corresponding to the χ^2 test of the order of the chain, that is, we are testing the null hypothesis that the order is k against the alternative that the order is $k + 1$.

Table 7: Order estimation for BP.

Order k	AIC(k)	BIC(k)	$\hat{h}_k$	$-2n(\hat{h}_{k+1} - \hat{h}_k)$	p-value	d.f.
$k = 0$	136.29	140.96	0.870	0.930	0.920	4
$k = 1$	143.36	157.35	0.864	10.107	0.607	12
$k = 2$	157.26	199.21	0.798			

Table 8: Order estimation for EP.

Order k	AIC(k)	BIC(k)	$\hat{h}_k$	$-2n(\hat{h}_{k+1} - \hat{h}_k)$	p-value	d.f.
$k = 0$	340.58	347.35	0.768	9.628	0.047	4
$k = 1$	338.95	359.28	0.746	12.667	0.394	12
$k = 2$	350.28	411.29	0.718			

¿From Table 7 it follows that all three criteria indicate that the order of the chain is $\ell = 0$ for BP. As for EP, Table 8 indicates that at a 5% level of significance the minimum entropy criterion and AIC coincide in $\ell = 1$. However the BIC indicates that the order should be zero.

Trying to understand the above discrepancy we performed a second set of analyses on samples of simulated data. We simulated four sequences of size 1,000 each, two for the BP and two for EP.

The first sequence was generated by a Markov chain of order zero with distribution given by Table 1. Table 9 shows the results for the estimation of the order for this sequence.

Table 9: Order estimation for the simulated data corresponding to Table 1 ($\ell = 0$).

Order k	AIC(k)	BIC(k)	$\hat{h}_k$	$-2n(\hat{h}_{k+1} - \hat{h}_k)$	p-value	d.f.
$k = 0$	1708.44	1718.26	0.853	1.476	0.831	4
$k = 1$	1714.97	1744.41	0.852	19.684	0.073	12
$k = 2$	1720.76	1809.08	0.843			

A second sequence was generated by a Markov chain of order one with transition probabilities given by Table 3. Table 10 shows the results for the estimation of the order for this sequence.

Table 10: Order estimation for the simulated data corresponding to Table 3 ($\ell = 1$).

Order k	AIC(k)	BIC(k)	$\hat{h}_k$	$-2n(\hat{h}_{k+1} - \hat{h}_k)$	p-value	d.f.
$k = 0$	1730.68	1740.49	0.864	7.951	0.093	4
$k = 1$	1730.72	1760.17	0.860	13.112	0.361	12
$k = 2$	1741.61	1829.93	0.854			

Making use of Tables 2 and 4, similar sequences were generated. Tables 11 and 12 contain the corresponding results.

Table 11: Order estimation for the simulated data corresponding to Table 2 ($\ell = 0$).

Order k	AIC(k)	BIC(k)	$\hat{h}_k$	$-2n(\hat{h}_{k+1} - \hat{h}_k)$	p-value	d.f.
$k = 0$	1488.51	1498.33	0.743	2.549	0.636	4
$k = 1$	1493.96	1523.40	0.742	13.551	0.330	12
$k = 2$	1504.41	1592.73	0.735			

Table 12: Order estimation for the simulated data corresponding to Table 4 ($\ell = 1$).

Order k	AIC(k)	BIC(k)	$\hat{h}_k$	$-2n(\hat{h}_{k+1} - \hat{h}_k)$	p-value	d.f.
$k = 0$	1785.59	1795.40	0.892	11.003	0.027	4
$k = 1$	1782.59	1812.03	0.886	20.089	0.065	12
$k = 2$	1786.50	1874.82	0.876			

Tables 9 to 12 show that the BIC criterion always indicates that the order of the chain should be zero, even though the simulated data comes from a Markov chain of order one. The AIC and the minimum entropy criterion however seem to be more precise indicating the right order for the chain at a 10% level of significance. This leads us to believe that in critical cases the AIC and the minimum entropy criterion are more reliable than the BIC.

¿From the above discussion, the AIC and the minimum entropy criterion indicate that BP and EP correspond to Markov chains of different orders. Since the BIC did not detect this difference, a test of goodness-of-fit was conducted assuming they were both indeed chains of order $\ell = 0$.

Table 13: Goodness-of-fit for EP assuming the law of BP ($\ell = 0$).

distances	1	2	3	n	U_n	p-value
observed values (EP)	32	160	27	219		
expected values	28.82	144.08	46.10	219	10.0280	0.0066

In the above table the statistics U_n is defined by (5), which tests the hypothesis that the sample from EP was produced by a sequence of independent random variables with probability distribution given by Table 1 (BP). The p-value indicates the rejection of the hypothesis meaning that BP and EP, if they were chain of order zero, they still have distinct distributions.

Acknowledgments We thank Maria Bernadete Abaurre, Pierre Collet, and Anthony Kroch for many helpful discussions and comments. This work was partially supported by CNPq, FAPESP 95/0790-1, PRONEX 41.96.0923.00, and USP 96.1.22073.1.6.

References

[1] Abaurre, M.B. et al, 1996, Data set presented at *3rd Workshop on Statistical Physics, Pattern Recognition and Grammar Selection*, IEA-USP.

[2] Anderson, T. W. and Goodman, L. A., 1957, Statistical Inference about Markov Chains, *Ann. Math. Statist.*, **28**, 89–110.

[3] Billingsley, P., 1961, Statistical Methods in Markov chains, *Annals of Mathematical Statistics*, **32**, 12-40.

[4] Cornfeld, I.P., Fomin, S.V., and Sinai, Ya.G., 1982, *Ergodic Theory*, Springer, New York.

[5] Frank, R., Galves, A., Galves, C. and Kroch, A., 1996, Prosodic patterns, parameter setting and language change, *IEA-USP, manuscript.*

[6] Katz, R.W., 1981, On some criteria for estimating the order of a Markov chain, *Technometrics*, **23**, 243-249.

[7] Kullback, S.L., 1959, *Information Theory and Statistics*, John Willey, New York.

[8] Nascimento, M. F. B. et al, 1987, *Português fundamental: métodos e documentos*, **2**, Instituto Nacional de Investigação Científica, Centro de Linguística da Universidade de Lisboa.

[9] Raftery, A. & Tavaré, S., 1994, Estimation and modeling repeated patterns in high order Markov chains with mixture transition distribution model, *Applied Statist. - J. of the Royal Statist. Soc., Series C*, **43**, 179-199.

[10] Schwarz, G., 1978, Estimating the dimension of a model, *The Annals of Statist.*, **6**, 461-464.

[11] Veilleux, N.M., Ostendorf, M., Price, P.J., and Shattuck Hufnagel, S., 1990, Markov modelling of prosodic phrase structure, in *International Conference on Acoustic, Speech, and Signal Processing - IEEE*, 777-780.

[12] Waterman, M.S. (ed), 1988, *Mathematical Methods for DNA Sequences*, Boca Raton.