Revista Colombiana de Estadística
Diciembre 2008, volumen 31, no. 2, pp. 229 a 240

Identificação de classes rítmicas de língua: modelagem de cadeias categorizadas da sonoridade usando árvores probabilísticas

Identifying Rhythmic Classes of Languages: Modeling Symbolics Chains of the Sonority Using Trees of Probability

Juvêncio Nobre[a]

Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará, Fortaleza, Brasil

Resumo

Recentemente, vários autores sugerem métodos para discriminar classes rítmicas de língua (Ramus et al. 1999, Duarte et al. 2001, Galves et al. 2002). Baseado no conceito de sonoridade, definido em Galves et al. (2002) e Cassandro et al. (2007), é proposto um modelo paramétrico para a família de processos estocásticos dos tempos de evolução da sonoridade para diferentes línguas, denotada por família de cadeias categorizadas ligadas. O objetivo do presente trabalho é modelar, para as diferentes línguas, as correspondentes cadeias categorizadas via cadeias de Markov de alcance variável (VLMC) e avaliar a conjectura de que estas resumem toda informação relevante dada pela sonoridade.

***Palavras chave*:** sonoridade, cadeias categorizadas ligadas, cadeias de Markov de alcance variável.

Abstract

Recently, several authors suggest methods to discriminate rhythmic classes of language (Ramus et al. 1999, Duarte et al. 2001, Galves et al. 2002). Based on sonority concept, defined in Galves et al. (2002), and Cassandro et al. (2007), a parametric model for the family of stochastic processes of sonority time evolution for different languages is proposed, denoted by family of tied quantized chains. The objective of this paper is to model, for the different languages, the correspondent quantized chains using Variable Length Markov Chains (VLMC) and evaluate the conjectures that summarize all relevant information given by the sonority.

***Key words*:** Sonority, Tied quantized chains, Variable length Markov chain.

[a]Professor adjunto I. E-mail: juvencio@ufc.br

1. Introdução e motivação

Desde meados do século passado, conjectura-se na literatura lingüística a existência de três classes rítmicas: línguas acentuais, silábicas e moraicas. Dado que nenhuma evidência a favor desta conjectura foi encontrada até metade dos anos de 1990, boa parte da comunidade científica questionava sobre a fundamentação da mesma. Mehler et al. (1996), baseados no fato de que recém-nascidos conseguiam discriminar grupos de frases de línguas de classes rítmicas conjecturadas distintas, forneceram evidências psico-lingüísticas da existência de classes rítmicas. Ramus et al. (1999) forneceram a primeira evidência favorável da existência das três classes conjecturadas.

A abordagem de Ramus et al. (1999) foi baseada em estatísticas descritivas do sinal acústico, proporção do tempo de duração em intervalos vocálicos ($\%V$) e desvio-padrão do tempo de duração nos intervalos consonantais (Δc). A escolha destas estatísticas baseiam-se em critérios lingüísticos. Foi analisado o sinal acústico de 20 sentenças produzidas por quatro falantes das seguintes línguas: inglês, polonês, holandês, catalão, espanhol, italiano, francês e japonês. Através de um gráfico de dispersão entre $\%V$ e Δc, perceberam, além de uma forte associação negativa entre as duas estatísticas, que as oito línguas em questão podem ser agrupadas em três classes: (i) inglês, polonês e holandês; (ii) espanhol, italiano, francês e catalão e (iii) japonês, que são por sua vez as três classes rítmicas conjecturadas na literatura, representando respectivamente as línguas *acentuais*, *silábicas* e *moraicas*. Esta abordagem é totalmente descritiva e apresenta alguns inconvenientes, principalmente no que tange a sua implementação (Galves et al. 2002).

Duarte et al. (2001) propõem um modelo paramétrico para o tempo de duração nos intervalos consonantais. Consideram que, para cada língua, estes tempos representam uma amostra aleatória de uma distribuição gama. A hipótese de interesse é que línguas pertencentes a uma mesma classe rítmica apresentam a mesma variância, a qual é diferente para línguas de outra classe rítmica. Considerando os dados utilizados em Ramus et al. (1999), e usando o teste da razão de verossimilhanças, obtiveram exatamente as mesmas classes. Galves et al. (2002) definiram um índice de regularidade do sinal de fala denotado por *sonoridade*. Tal índice é uma função do sinal acústico evoluindo dentro do intervalo $[0, 1]$, de tal forma que assume valores próximos a 1 quando a região é *regular* (regiões sonoras), e valores próximos a 0 quando a região é *irregular* (regiões com muita obstrução). Os autores sugerem que, através de uma análise das trajetórias da sonoridade de cada língua, é possível discriminar as três classes rítmicas existentes.

Recentemente, Cassandro et al. (2007) propõem um modelo paramétrico para a família de processos estocásticos dos tempos de evolução da sonoridade para diferentes línguas, denotada por família de cadeias categorizadas ligadas. As cadeias são ditas *ligadas* pela suposição de existência de uma partição universal do domínio da sonoridade, de tal forma que a distribuição da sonoridade, condicionada em cada intervalo (definidos pela partição) independe da língua. Um procedimento para estimar os pontos de corte (definidores da partição) de forma consistente também é apresentado. A seguinte frase de Cuesta-Albertos et al. (2007): "*... the most important linguistic question of the existence of rhytmic classe should be*

decided using only the properties of the symbolics chains" motivou o presente trabalho, que tem por objetivo avaliar a conjectura de que as cadeias categorizadas resumem toda a informação relevante dada pela sonoridade e que estas podem ser utilizadas para discriminar as classes rítmicas de fala.

As cadeias categorizadas serão modeladas via cadeias de Markov de alcance variável. Os dados utilizados neste trabalho são descritos na Seção 2, enquanto que a metodologia é descrita na Seção 3. Na Seção 4 apresentam-se os resultados da análise, e na Seção 5 são discutidos os resultados obtidos.

2. Conjunto de dados

Neste trabalho são utilizados os dados lingüísticos analisados em Ramus et al. (1999), que constituem 160 sentenças de 8 diferentes línguas: inglês, polonês, holandês, catalão, espanhol, italiano, francês e japonês. Para cada língua, foram utilizadas 20 sentenças, selecionadas de um total de 54 com objetivo de controlá-las com relação ao número de sílabas, produzidas por 4 mulheres. A justificativa para tal seleção é eliminar possíveis sentenças *discrepantes* com respeito à média do tempo de fala. Uma análise de tais seqüências, usando a sonoridade, é apresentada em Galves et al. (2002).

3. Modelagem

Nas duas subseções a seguir, descrevemos a modelagem probabilística utilizada para modelar o conjunto de dados em questão.

3.1. Cadeias de Markov de alcance variável

As cadeias de Markov representam uma boa alternativa para modelar estrutura de dependência, por exemplo, para aplicações em estudos com medidas repetidas (Ware et al. 1988, Reboussin 1990, Chao & Kosorok 1995, Lindsey 1999, Agresti 2002). Porém, do ponto de vista estatístico, esses modelos não são muito atraentes, dado o número elevado de parâmetros a serem estimados em certas circunstâncias. Como ilustração, considere $\{X_n\}_{n\in\mathbb{N}}$ uma cadeia de Markov de ordem k, definida em um alfabeto finito A. O número de parâmetros não redundantes[1] a serem estimados é $|A|^k(|A|-1) = O(|A|^{k+1})$; $|A|$ representa a cardinalidade de A. Perceba que o número de parâmetros a serem estimados cresce exponencialmente em k; por conseguinte, para valores moderamente elevados de k, tem-se um número muito grande de parâmetros a serem estimados.

Uma classe de modelos mais parcimoniosa, Com relação ao número de parâmetros a serem estimados, são as cadeias de Markov de alcance variável (VLMC, do inglês *Variable Length Markov Chains*), cuja gênese é devida a Rissanen (1983),

[1]Linearmente independentes.

no contexto de teoria da informação, e foi recentemente discutida e popularizada por Bühlmann & Wyner (1999) dentro do ponto de vista estatístico.

Considere $\{X_n\}_{n\in\mathbb{N}}$ uma cadeia de Markov estacionária definida em um alfabeto finito A. Denotando, $\{w \in \Omega \mid X_{-1}(w) = x_{-1}, \ldots, X_{-k}(w) = x_{-k}\} := x_{-k}^{-1}$ e $\mathbb{P}(X_0 = x_0 \mid X_{-1} = x_{-1}, \ldots, X_{-k} = x_{-k}) = \mathbb{P}(x_0 \mid x_{-k}^{-1})$, $\forall k \in \mathbb{N}$, a função

$$\begin{aligned} c : A^{\infty} &\longrightarrow \bigcup_{j=0}^{\infty} A^j \\ x_{-\infty}^{0} &\longmapsto x_{-l}^{0} \end{aligned}$$

em que $l = l\big(x_{-\infty}^{-1}\big) := \min\big\{k \mid \mathbb{P}\big(x_0 \mid x_{-\infty}^{-1}\big) = \mathbb{P}\big(x_0 \mid x_{-k}^{-1}\big), \forall x_0 \in A\big\}$, é denotada como *função contexto* da cadeia. A denominação *contexto* é devida a que apenas parte do passado é *relevante* para a variável X_0, e esta é uma função da configuração $x_{-\infty}^{-1}$. A função l indica a quantidade de passados relevantes. Denotando $w < \infty$ o menor inteiro tal que $\big|c\big(x_{-\infty}^{-1}\big)\big| = l \leq w, \forall x_{-\infty}^{-1} \in A^{\infty}$, então $\{X_n\}_{n\in\mathbb{N}}$ é dita ser uma cadeia de Markov de alcance variável de ordem w. A forma mais conveniente de representar esta classe de modelos é através da sua árvore de contexto. Para uma cadeia de Markov de alcance variável $\{X_n\}_{n\in\mathbb{N}}$ de ordem w com função contexto c, sua árvore de contexto é definida por uma árvore com ramos $\big\{s \mid s = c\big(x_{-\infty}^{-1}\big), \forall x_{-\infty}^{-1} \in A^{\infty}\big\}$.

É interessante perceber que a função contexto pode ser obtida diretamente da árvore de contexto que nada mais é que o conjunto dos passados relevantes (seus ramos). A árvore do contexto pode ser caracterizada da seguinte forma:

1. O primeiro nó é a raiz, enquanto que os nós das extremidades inferiores são chamados terminais.

2. Os galhos representam os passados relevantes (do mais próximo ao mais longínquo).

3. Cada nó tem no máximo $|A|$ arestas.

4. O contexto é representado pelos galhos que ligam o primeiro e último nó.

Uma cadeia $\{X_n\}_{n\in\mathbb{N}}$ (VLMC) é completamente determinada por sua árvore de contextos. Para maiores detalhes sobre cadeias de Markov de alcance variável e sua representação através de árvores de contexto, veja Bühlmann & Wyner (1999) e Ferrari & Wyner (2003), por exemplo.

O processo de estimação, baseado em uma amostra observada, da matriz de transição de uma cadeia de Markov de alcance variável de ordem w é feito via *algoritmo do contexto* proposto por Rissanen (1983). Para detalhes e comentários sob a consistência do método, veja por exemplo Rissanen (1983), Bühlmann & Wyner (1999) e Bühlmann (2000). Para discussão a respeito de seleção de modelos, veja Bühlmann (2000); Mäechler & Bühlmann (2004) apresentam um tutorial sobre o pacote VLMC desenvolvido em linguagem R (R Development Core Team 2007) para ajustes destes tipos de modelos.

3.2. Cadeias categorizadas ligadas

A noção de cadeias categorizadas ligadas teve sua gênese no trabalho de Cassandro et al. (2007), que a propõem como um modelo paramétrico para a família de processos estocásticos dos tempos de evolução da sonoridade para as diferentes línguas. Tal família é descrita sucintamente a seguir.

Para cada língua $l \in L = \{1, \ldots, 8\}$ considera-se um processo estocástico $\left\{\left(S_t^l\right)_{t\in\mathbb{N}}\ l \in L\right\}$ assumindo valores no intervalo $[0, 1]$ representando os tempos de evolução da sonoridade para a l-ésima língua. Assume-se que os processos supracitados são *estacionários* e *ergódicos*. Tais processos são *ligados* sob a suposição de existência de um número inteiro N e de uma seqüência crescente de pontos de corte

$$0 = c_0 < c_1 < \cdots < c_N < c_{N+1} = 1$$

e $N + 1$ medidas de probabilidade π_j $(j = 0, \ldots, N)$ com respectivo suporte $I_j = (c_j, c_{j+1}]$, tal que $\forall t \in \mathbb{N}$ e $\forall l \in L$

$$\mathbb{P}[S_t^l \in B \mid S_t^l \in I_j] = \pi_j(B) \tag{1}$$

com B representando um boreliano do intervalo $[0, 1]$.

Por hipótese, os pontos de corte c_j e as medidas de probabilidade π_j, $j = 0, \ldots, N$ são universais, isto é, independem de l. Os intervalos I_j constituem regiões com diferentes níveis de sonoridade. A cadeia categorizada $\{X_t^l\}_{t\in\mathbb{N}}$, que assume valores no alfabeto finito $A = \{0, \ldots, N\}$, é definida por

$$X_t^l = \sum_{j=0}^{N} j\mathbb{1}(S_t^l \in I_j), \qquad \forall t \in \mathbb{N}$$

Sob as suposições a respeito de $\left\{S_t^l\right\}_{t\in\mathbb{N}}$, as cadeias categorizadas $\left\{X_t^l\right\}_{t\in\mathbb{N}}$ também são estacionárias e ergódicas. No presente trabalho, a categorização foi feita utilizando os quatro pontos de cortes universais estimados de forma consistente em Cassandro et al. (2007), $c_1 = 0.19$, $c_2 = 0.46$, $c_3 = 0.67$ e $c_4 = 0.93$, através do método *bootstrap*.

4. Análise estatística

Para cada sentença, 20 em cada língua, foi obtida a sonoridade e, posteriormente, a cadeia categorizada associada. As cadeias categorizadas foram modeladas via cadeias de Markov de alcance variável; desta forma, obteve-se uma árvore de contexto estimada para cada cadeia categorizada (sentença).

Para efeito de ajuste, foi utilizado o pacote `VLMC` (Mäechler 2006), desenvolvido em linguagem R. Foi utilizada a correção de Bonferroni, obtendo assim uma aproximação para o ponto de corte usado no algoritmo do contexto que depende do tamanho da seqüência, conforme é sugerido em Bühlmann (2000). Na figura 1 mostram-se as árvores de contextos estimadas mais freqüentes, as referentes cadeias, que serão doravante denominadas por cadeia 1, 2 e 3, respectivamente. As

demais cadeias estimadas que foram observadas apresentam uma freqüência muito baixa; no máximo em duas sentenças para cada língua. Desta forma, para efeito de análise, foram consideradas apenas estas três cadeias supracitadas. Na tabela 1, mostram-se o percentual das sentenças (%) para cada língua, no qual observa-se a referida árvore de contexto estimada.

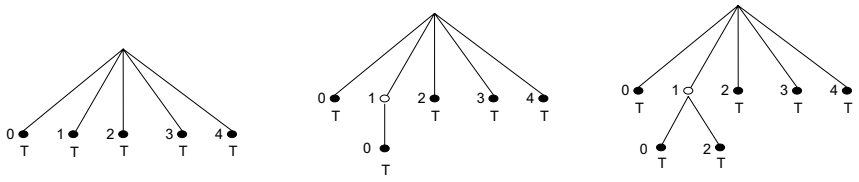

FIGURA 1: Cadeias estimadas mais freqüentes.

TABELA 1: Proporção de cadeias estimadas para cada língua.

Língua	Cadeia 1	Cadeia 2	Cadeia 3
Japonês	25%	35%	$\leq 5\%$
Polonês	35%	45%	$\leq 5\%$
Holandês	30%	30%	15%
Inglês	15%	15%	30%
Espanhol	35%	30%	$\leq 5\%$
Francês	30%	15%	$\leq 5\%$
Italiano	30%	20%	$\leq 5\%$
Catalão	40%	$\leq 5\%$	$\leq 5\%$

Para cada língua, foi considerada a matriz de probabilidades de transição estimada da seqüencia que apresentava o menor BIC (*Bayesian Information Criterion*) para cada uma das duas (ou três, para o holandês e inglês) cadeias (as probabilidades de transição estimadas são mostradas no Apêndice A.). Para as línguas ditas como *silábicas*, com exceção do **catalão**, na maior parte das sentenças, a cadeia 1 traduz melhor o comportamento da sentença, precedida de uma porcentagem um pouco menor do número de sentenças no qual a cadeia 2 foi a que melhor modelou seu comportamento. Para as línguas ditas *acentuais*, com exceção do **polonês**, não existe nenhuma dominação das três cadeias. A proporção de sentenças que apresentaram a cadeia 1 como a "cadeia verdadeira" é igual a proporção de sentenças que apresentaram a cadeia 2 como a "verdadeira". Com relação ao japonês, existe uma ligeira "dominação" A favor da cadeia 2; no entanto, esta pode não se confirmar caso se analise um número maior de sentenças. Com relação ao polonês e ao catalão, essas duas línguas não se comportam de forma similar às demais línguas pertencentes as suas respectivas classes rítmicas. O polonês apresentou comportamento similar ao japonês, porém menor variabilidade no que tange as árvores estimadas.

Desta forma, considerando apenas as línguas holandês, inglês, francês, italiano, japonês e espanhol, e os aspectos acima mencionados, há evidências favoráveis a existência de três clusters: o primeiro, formado pelas línguas holandês e inglês, caracteriza as línguas *acentuais*; o segundo, formado pelas línguas francês, italiano e espanhol, caracteriza as línguas *silábicas*; e o terceiro, formado apenas pelo japonês. Tal resultado é análogo ao resultado obtido em Cuesta-Albertos et al. (2007) usando a sonoridade, em que tanto o polonês como o catalão não foram discriminados, sendo o mesmo compatível com a conjectura lingüística de existência de três classes rítmicas.

5. Resultados e discussão

Pelos resultados mostrados na seção anterior, pode concluir-se que as cadeias categorizadas resumem toda a informação relevante dada pela sonoridade; conseqüentemente, podem ser utilizadas para discriminar as classes rítmicas de fala.

Um tópico interessante a ser pesquisado seria avaliar o ajuste de modelos do tipo **mistura** de cadeias de Markov de alcance variável, em que se poderia considerar, para as classes de línguas *moraicas* e *silábicas*, que cada sentença seria modelada por uma mistura das cadeias 1 e 2; o parâmetro de mistura possivelmente seja diferente para as duas classes, enquanto que para as línguas *acentuais*, cada sentença seria modelada por uma mistura das cadeias 1, 2 e 3.

Agradecimentos

Este trabalho foi desenvolvido e apresentado durante a disciplina MAE 5741 -Inferência em Processos Estocásticos, ministrada pelo professor Antônio Galves no IME-USP em 2005. O autor gostaria de agradecer ao professor Antônio Galves e a Anne Cros e aos dois árbitros que concederam imprescindíveis sugestões para o melhoramento deste trabalho.

[Recibido: febrero de 2008 — Aceptado: octubre de 2008]

Referências

Agresti, A. (2002), *Categorical Data Analysis*, John Wiley & Sons, New York, United States.

Bühlmann, P. (2000), 'Model Selection for Variable Length Markov Chains and Tuning the Context Algorithm', *Ann. Inst. Statist. Math.* **25**, 287–315.

Bühlmann, P. & Wyner, A. J. (1999), 'Variable Length Markov Chains', *Annals of Statistics* **27**, 480–513.

Cassandro, M., Collet, P., Duarte, D., Galves, A. & Garcia, J. (2007), 'A stochastic Model for the Speech Sonority: Tied Quantized Chains and Cross-Linguistic Estimation of the Cut-Points', *Math. & Sci. hum.* **180**, 43–55. Mathematical Social Sciences, 45 année.

Chao, W. H. & Kosorok, M. R. (1995), Asymptotic Properties of Markov Regression Models for Longitudinal Categorical Data in Continuous Time, Biostatistic technical report, Department of Statistic, University of Wisconsin.

Cuesta-Albertos, J., Fraiman, R., Galves, A. & Garcia, J. (2007), 'Identifying Rhythmic Classes of Languages Using their Sonority: A Kolmogorov-Smirnov Approach', *Journal of Applied Statistics* **34**, 749–761.

Duarte, D., Galves, A., Lopes, N. & Maronna, R. (2001), Robust Test for Equality of Variances the Statistical Analysis of Acoustic Correlates of Speech Rhythm, *in* 'Parameter setting and language change', Workshop on rhythmic patterns, University of Bielefeld.
*http://www.physik.uni-bielefeld.de/complexity/duarte.pdf

Ferrari, F. & Wyner, A. (2003), 'Estimation of General Stationary Processe by Variable Length Markov Chains', *Scandinavian Journal of Statistics* **30**, 459–480.

Galves, A., Garcia, J., Duarte, D. & Galves, C. (2002), Sonority as a Basis for Rhythmic Class Discrimination, *in* 'Speech Prosody'.
*www.lpl.uinv-aix.fr/sp2002/pdf/galves-et-al.pdf

Lindsey, J. K. (1999), *Models for Repeated Measurements*, second edn, Oxford Statistical series, New York, United States.

Mäechler, M. (2006), *VLMC: VLMC–Variable Length Markov Chains*. R package version 1.3-10.

Mäechler, M. & Bühlmann, P. (2004), 'Variable Length Markov Chains: Methodology, Computing and Software', *Journal of Computational & Graphical Statistics* **13**, 435–455.

Mehler, J., Dupoux, E., Nazzi, T. & Dehaene-Lambertz, G. (1996), Coping with Linguistic Diversity: The Infant's Viewpoint, *in* J. L. Morgan & K. Demuth, eds, 'Signal to syntax: bootstraping from speech to grammar in early acquisition'.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Ramus, F., Nespor, M. & Mehler, J. (1999), 'Correlates of Linguistic Rhythm in the Speech Signal', *Cognition* **73**, 265–292.

Reboussin, D. V. (1990), Discovering Markov Structure in Group Sequential Methods for Longitudinal Studies, Biostatistic technical report 61, Department of Statistic, University of Wisconsin.

Rissanen, J. (1983), 'A Universal Data Compression System', *IEEE Trans. Inform. Theory* **29**, 656–664.

Ware, J., Lipsitz, S. & Speizer, F. (1988), 'Issues in the Analysis of Repeated Categorical Outcomes', *Statistics in Medicine* **7**, 95–107.

Apêndice A.

A seguir, apresentamos as matrizes de transição estimadas das sentenças que apresentaram menor BIC, para as cadeias mais freqüentes, para as oito línguas.

Tabela 2: Matriz de transição estimada da cadeia 1 para o japonês.

Contexto	0	1	2	3	4
0	0.69	0.27	0.04	0.00	0.00
1	0.16	0.57	0.23	0.04	0.00
2	0.01	0.21	0.60	0.18	0.00
3	0.00	0.00	0.08	0.82	0.10
4	0.00	0.00	0.00	0.10	0.90

Tabela 3: Matriz de transição estimada da cadeia 2 para o japonês.

Contexto	0	1	2	3	4
0	0.67	0.28	0.06	0.00	0.00
$[x; x \neq 0]1$	0.16	0.51	0.29	0.03	0.00
01	0.00	0.24	0.64	0.12	0.00
2	0.03	0.26	0.45	0.26	0.00
3	0.00	0.01	0.11	0.82	0.06
4	0.00	0.00	0.00	0.08	0.92

Tabela 4: Matriz de transição estimada da cadeia 1 para o polonês.

Contexto	0	1	2	3	4
0	0.65	0.32	0.01	0.02	0.00
1	0.18	0.53	0.23	0.06	0.00
2	0.03	0.29	0.44	0.24	0.00
3	0.00	0.02	0.09	0.80	0.09
4	0.00	0.00	0.00	0.07	0.93

Tabela 5: Matriz de transição estimada da cadeia 2 para o polonês.

Contexto	0	1	2	3	4
0	0.69	0.26	0.05	0.00	0.00
$[x.x \neq 0]1$	0.16	0.56	0.26	0.02	0.00
01	0.00	0.39	0.54	0.07	0.00
2	0.03	0.26	0.52	0.20	0.00
3	0.00	0.02	0.07	0.82	0.09
4	0.00	0.00	0.00	0.08	0.92

Tabela 6: Matriz de transição estimada da cadeia 1 para o holandês.

Contexto	0	1	2	3	4
0	0.69	0.27	0.04	0.00	0.00
1	0.23	0.51	0.20	0.06	0.00
2	0.02	0.21	0.51	0.26	0.00
3	0.00	0.01	0.11	0.79	0.09
4	0.00	0.00	0.00	0.07	0.93

Tabela 7: Matriz de transição estimada da cadeia 2 para o holandês.

Contexto	0	1	2	3	4
0	0.64	0.35	0.01	0.00	0.00
$[x; x \neq 0]1$	0.21	0.52	0.21	0.06	0.00
01	0.04	0.35	0.43	0.18	0.00
2	0.02	0.31	0.51	0.16	0.00
3	0.00	0.01	0.09	0.80	0.10
4	0.00	0.00	0.00	0.12	0.88

Tabela 8: Matriz de transição estimada da cadeia 3 para o holandês.

Contexto	0	1	2	3	4
0	0.64	0.33	0.02	0.01	0.00
$\big[x; x \in \{1,3,4\}\big]1$	0.20	0.55	0.22	0.03	0.00
01	0.06	0.46	0.42	0.06	0.00
21	0.35	0.58	0.05	0.02	0.00
2	0.01	0.30	0.50	0.19	0.00
3	0.00	0.02	0.10	0.83	0.05
4	0.00	0.00	0.00	0.09	0.91

Tabela 9: Matriz de transição estimada da cadeia 1 para o inglês.

Contexto	0	1	2	3	4
0	0.65	0.30	0.05	0.00	0.00
1	0.16	0.59	0.21	0.04	0.00
2	0.02	0.24	0.53	0.21	0.00
3	0.00	0.01	0.11	0.79	0.09
4	0.00	0.00	0.00	0.05	0.95

Tabela 10: Matriz de transição estimada da cadeia 2 para o inglês.

Contexto	0	1	2	3	4
0	0.62	0.32	0.06	0.00	0.00
$[x; x \neq 0]1$	0.23	0.47	0.26	0.04	0.00
01	0.02	0.40	0.47	0.11	0.00
2	0.03	0.30	0.46	0.21	0.00
3	0.00	0.02	0.09	0.80	0.09
4	0.00	0.00	0.00	0.07	0.93

Tabela 11: Matriz de transição estimada da cadeia 3 para o inglês.

Contexto	0	1	2	3	4
0	0.64	0.33	0.03	0.00	0.0
$\big[x; x \in \{1, 3, 4\}\big]1$	0.21	0.50	0.26	0.03	0.00
01	0.02	0.31	0.53	0.13	0.00
21	0.35	0.60	0.06	0.00	0.00
2	0.02	0.23	0.55	0.20	0.00
3	0.00	0.02	0.12	0.80	0.10
4	0.00	0.00	0.00	0.12	0.88

Tabela 12: Matriz de transição estimada da cadeia 1 para o espanhol.

Contexto	0	1	2	3	4
0	0.68	0.31	0.01	0.00	0.00
1	0.14	0.58	0.24	0.04	0.00
2	0.01	0.26	0.52	0.21	0.00
3	0.00	0.01	0.09	0.80	0.10
4	0.00	0.00	0.00	0.08	0.92

Tabela 13: Matriz de transição estimada da cadeia 2 para o espanhol.

Contexto	0	1	2	3	4
0	0.57	0.36	0.07	0.00	0.00
$[x; x \neq 0]1$	0.19	0.56	0.25	0.00	0.00
01	0.04	0.32	0.64	0.00	0.00
2	0.01	0.27	0.52	0.20	0.00
3	0.00	0.01	0.05	0.84	0.10
4	0.00	0.00	0.00	0.06	0.94

Tabela 14: Matriz de transição estimada da cadeia 1 para o francês.

Contexto	0	1	2	3	4
0	0.65	0.29	0.06	0.00	0.00
1	0.17	0.52	0.27	0.04	0.00
2	0.01	0.30	0.46	0.23	0.00
3	0.00	0.01	0.05	0.85	0.09
4	0.00	0.00	0.00	0.09	0.91

Tabela 15: Matriz de transição estimada da cadeia 2 para o francês.

Contexto	0	1	2	3	4
0	0.65	0.25	0.10	0.00	0.00
$[x; x \neq 0]1$	0.21	0.47	0.29	0.03	0.00
01	0.05	0.24	0.71	0.00	0.00
2	0.00	0.27	0.46	0.27	0.00
3	0.00	0.01	0.10	0.81	0.08
4	0.00	0.00	0.00	0.10	0.90

Tabela 16: Matriz de transição estimada da cadeia 1 para o italiano.

Contexto	0	1	2	3	4
0	0.67	0.28	0.05	0.00	0.00
1	0.16	0.59	0.22	0.03	0.00
2	0.01	0.25	0.43	0.31	0.00
3	0.00	0.00	0.11	0.82	0.07
4	0.00	0.00	0.00	0.07	0.93

Tabela 17: Matriz de transição estimada da cadeia 2 para o italiano.

Contexto	0	1	2	3	4
0	0.70	0.26	0.04	0.00	0.00
$[x; x \neq 0]1$	0.22	0.51	0.20	0.07	0.00
01	0.06	0.36	0.33	0.25	0.00
2	0.03	0.27	0.46	0.24	0.00
3	0.00	0.01	0.09	0.81	0.09
4	0.00	0.00	0.00	0.11	0.89

Tabela 18: Matriz de transição estimada da cadeia 1 para o catalão.

Contexto	0	1	2	3	4
0	0.68	0.29	0.03	0.00	0.00
1	0.13	0.57	0.28	0.02	0.00
2	0.02	0.15	0.61	0.22	0.00
3	0.00	0.01	0.09	0.82	0.08
4	0.00	0.00	0.00	0.11	0.89