Correlates of rhythm in written texts of Brazilian and European Portuguese

Antonio Galves[*], Charlotte Galves[†], Nancy Garcia[‡], Cláudia Peixoto[§]

April 22, 2008

Abstract

We address the question of detecting fingerprints of rhythm in written texts. We study texts from 20th century Brazilian and Portuguese authors. In these texts we codify the syllables according to whether they carry main stress or not, and whether they are at the beginning of a phonological word or not. Additionally, periods are also marked. Modeling the sequences of symbols obtained with this codification as Variable Length Markov Chains, we estimate the patterns for each text. This probabilistic model discriminates European Portuguese from Brazilian Portuguese. Moreover, the model obtained for each language has a clear linguistic interpretation, as it captures structural features which have long been conjectured in the literature.

1 Introduction

This paper addresses the question of the existence of fingerprints of rhythm in written texts. The data we analyze is composed by codified written texts from Portuguese and Brazilian 20th century authors. Codification expresses only a few basic rhythmic features which can be retrieved from written texts, with no extra information concerning the way the texts would be effectively read by native speakers.

Modern Portuguese provides an interesting case to be analyzed from the point of view of rhythm. European Portuguese and Brazilian Portuguese (henceforth EP and BP respectively) have the same *lexicon* . However they have been conjectured to implement different *rhythms* (cf. for instance Carvalho 1989, Frota and Vigário 2000; 2001, Sândalo *et al.* to appear and Duarte *et al.* 2001).

This last paper propose a parametric family of probability distributions that closely fits the data in Ramus *et al.*(1999). This makes it possible to perform

[*]Universidade de São Paulo, São Paulo, Brasil, galves@ime.usp.br
[†]Universidade Estadual de Campinas, Campinas, Brasil, galvesc@unicamp.br
[‡]Universidade Estadual de Campinas, Campinas, Brasil, nancy@ime.unicamp.br
[§]Universidade de São Paulo, São Paulo, Brasil, claudiap@ime.usp.br

statistical inference, i. e., to extend results from the sample (the data set) to the population (the set of all potential sentences). This model enables testing the hypothesis that EP is stress-timed language, while BP is syllable-timed language . It turns out that the speech data from BP and EP speakers analyzed in Duarte *et al.* 2001 is compatible with the conjecture.

The question we address here is whether it is possible to detect this difference in written texts. In the absence of phonetic implementation we tried to follow the trail suggested by the conjecture that rhythmic classes are also characterized by the fact that they assign relevance to different prosodic domains. The questions is how to retrieve on statistical grounds relevant domains out from a sample of codified written texts. *Variable Length Markov Chains (VLMC)* was a natural tool to consider in order to perform the relevant statistical analysis.

This class of models were first introduced in the information theory literature by Rissanen (1983) under the name of *finite memory source* or *probabilistic tree*. More recently this class of models have become quite popular in the statistics literature under the name of *Variable Length Markov Chains (VLMC)* (Bühlman and Wyner, 1999)

VLMC is a flexible class of Markov chains in which the part of the past which is relevant to predict the next symbol has a variable length depending on the observed past values. These relevant parts of the past are called *contexts*.

The notion of context makes VLMC models parsimonious, with less parameters to estimate than the traditional approach to Markov chains in which the number of parameters grows exponentially with the length of the fixed past.

VLMC models are particularly interesting in phonology, as left boundaries of contexts can be interpreted as left boundaries of relevant prosodic domains. This last observation is at the origin of the present work. It was tempting to check if a blind, plain statistical analysis, without any *a priori* assumption, would be able to detect boundaries relevant for the implementation of rhythmic patterns in one language but not in the other. The result went beyond our expectations.

There is a consistent statistical procedure to identify the set of relevant contexts out from a sample. This is the so called *Context Algorithm*. Applied to a sample of BP and EP codified written texts this procedure puts in evidence a striking difference in the set of contexts describing BP and EP. It turns out that this difference has a clear linguistic interpretation and corresponds to a feature which has been conjectured in the literature concerning the rhythmic differences between BP and EP.

2 The data

The data we analyze is a codified corpus of written texts of Brazilian and Portuguese 20^{th} century writers. Codification was made by assigning two symbols to each syllable of the text according to whether

- it is stressed or not;

- it is the beginning of a prosodic word or not,

where by *prosodic word* we mean a lexical word together with the functional non stressed words which precede it (cf. for instance Vigário 2003).

This amounts to use $\{0,1\}^2$ as the set of symbols where the first symbol indicates if the syllable is the beginning or not of a prosodic word and the second symbol indicates if the syllable is stressed or not. To simplify the notation we will use the binary expansion to represent the pairs as integers as follows $(0,0) = 0$, $(0,1) = 1$, $(1,0) = 2$ and $(1,1) = 3$. Additionally we will assign an extra symbol (4) to codify the the beginning of each sentence. Let us call $\mathcal{A} = \{0,1,2,3,4\}$ the alphabet obtained in this way.

An example will help understanding the codification. The sentence *O menino já comeu o doce* (the boy already ate the candy) starts with the prosodic word *O menino* which is codified as

Sentence	.	O	me	ni	no
begining of a prosodic word		1	0	0	0
stressed symbol		0	0	1	0
Code	4	2	0	1	0

The binary digits 0 and 1 indicates if the condition *"beginning of a prosodic word"* or *"stressed symbol"* are satisfied or not. The last line presents the pair of binary symbols using the binary expansion. The opening symbol 4 indicates that the sentence starts with this prosodic word. The total sentence is codified as

.	O	me	ni	no	já	co	meu	o	do	ce
4	2	0	1	0	3	2	1	2	1	0

It is worth observing that Portuguese morphology imposes a set of constraints on the symbolic chains produced by the codification of the texts. First of all, prosodic words are finite and their lengths are bounded above by a fixed constant M. For practical purposes, we can take $M = 15$. Moreover according to the definition we adopted, any prosodic word must contain one and only one stressed syllable (codified by 1 or 3). Furthermore, Portuguese allows a stressed syllable (1 or 3) to be followed by at most three unstressed (codified 0) syllables inside a prosodic word. Finally, as sentences are formed by the concatenation of prosodic words, at the beginning of a sentence (codified by symbol 4) the only symbols which are allowed are the symbols associated to the beginning of a prosodic word (2 or 3).

This codification can be performed automatically. In our case we used a software written in Perl which can be freely downloaded for academic purposes at URL `www.ime.usp.br/~tycho/prosody/vlmc/tools/silaba.pl`.

3 The probabilistic model

We will model each codified text as a finite sample of a stationary Variable Length Markov Chain (VLMC). $(X_i)_{i\in\mathbb{Z}}$ with values on the alphabet $\mathcal{A} =$

$\{0,1,2,3,4\}$. We recall that a VLMC is a Markov chain of finite order whose probability transtions has the following property

$$\mathbb{P}(X_0 = x_0 \mid X_{-K}^{-1} = x_{-K}^{-1}) =$$

$$\mathbb{P}(X_0 = x_0 \mid X_{-K}^{-1} = x_{-\ell(x_{-K}^{-1})}^{-1}) \tag{1}$$

where $\ell : \mathcal{A}^K \to \{1, \ldots, K\}$ is a function of the past which indicates the number of steps we must look back in each past to choose the next outcome of the chain.

In the above definition we used the short notation

$$\mathbb{P}(X_0 = x_0 \mid X_{-1} = x_{-1}, \ldots, X_{-K} = x_{-K}) =$$

$$\mathbb{P}(X_0 = x_0 \mid X_{-K}^{-1} = x_{-K}^{-1})\,.$$

The function $c : \mathcal{A}^K \to \cup_{m=1}^K \mathcal{A}^m$

$$c : x_{-K}^{-1} \to x_{-\ell(x_{-K}^{-1})}^{-1} \tag{2}$$

is called the *context function* of the chain. We observe that the context function satisfies the *suffix property*, which means that no context is a suffix of another context. Due to this property, a context function $c(\cdot) : \mathcal{A}^K \to \cup_{m=1}^K \mathcal{A}^m$ can be represented as a tree constructed as follows

- Root on the top;
- Branches grow downwards;
- Every internal node has at most $|\mathcal{A}|$ offsprings;
- The context $w = c(x_{-K}^{-1})$ is represented by a branch, whose sub-branch on top is determined by x_{-1}, the next sub-branch is determined by x_{-2} and so on.
- For each branch this procedure is repeated $\ell(x_{-K}^{-1})$ times.

A VLMC is completely determined by its tree of contexts and by the set of probability transitions associated with each branch of the tree.

Given a large sample produced by a VLMC we can estimate the tree of contexts of the VLMC as well as its associated set of probability transitions using the *Context Algorithm*. This is a consistent algorithm which was introduced by Rissanen 1983 and re-proposed by Bühlmann and Wyner (1999). The way the algorithm works can be summarized as follows. We start with a maximal tree of candidate contexts. This maximal tree is constructed by considering only branches (*i.e.* sequences of symbols) which appear at least a fixed number of times in the sample. Given a candidate tree of contexts we associate to each context a probability transition which is estimate from the sample by a maximum likelihood procedure, (*i.e.* given a context we estimate the probability of

a given symbol to appear in the next position as the proportion of times in the sample the context is followed by this symbol). The estimation algorithm is performed by pruning sequentially the initial maximal tree of candidate contexts. The estimated context tree is the smallest tree which maximizes the likelihood of the sample. The estimated set of probability transitions is the one associated to that by the maximum likelihood procedure.

4 Statistical analysis

We analyzed a set of codified texts of 20th century BP authors (35 texts) and EP authors (13 texts). To analyze the data we used the VLMC package, which is GNU free software developed by Martin Mächeler as part of the *R-project* (http://cran.r-project.org). For a tutorial on the VLMC package we refer the reader to Mächler (2005) and to Mächler and Bühlmann (2004).

Figure 2 shows the typical tree of contexts estimated from a codified sample produced by an BP author and Figure 1 shows the typical tree of contexts estimated from a codified sample produced by an EP author.

Both trees share the property that the strings $(2), (3)$ and (4) are contexts. This means that every time the last symbol in the past is $2, 3$ or 4 the prediction of the next symbol can be performed without extra knowledge of the past.

Both trees also share the fact that the symbol 0 alone is not enough to predict the past. Both in BP and in EP when the last symbol in the past is a 0 we must look back two extra steps in the past before being able to predict the next symbol.

These common facts can be summarized as follows: both in BP and in EP every time the chain visits the set $\{2, 3, 4\}$ it forgets the past.

The absence in both trees of several strings of symbols is explained by constraints imposed by the morphology of Portuguese which makes impossible the existence of these strings.

The crucial difference between the two trees is the role played by the symbol 1 which defines a context in EP, but not in EP. This is main finding of the analysis.

The probabilities transitions associated to each tree are presented bellow. Their interpretation is less clear.

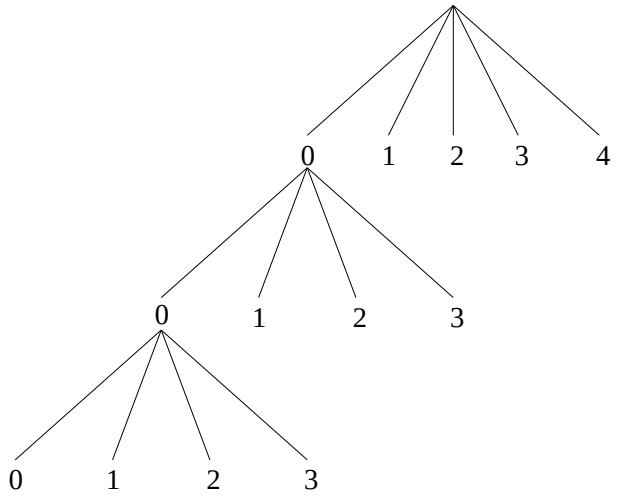

Figure 1: Example of a EP tree.

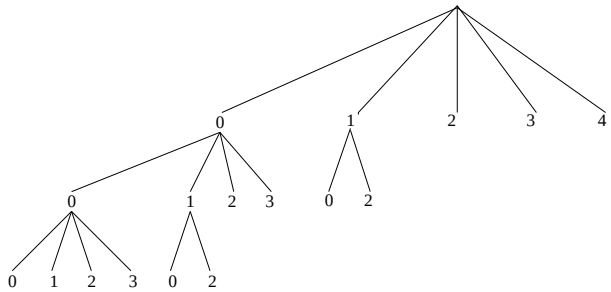

Figure 2: Example of a BP tree

Contexts	Probability							
	0	1	2	3	4	n	freq	
000	0,31	0,69	0	0	0	238	0,035	
100	0	0	0,80	0,14	0,06	103	0,015	
200	0,37	0,63	0	0	0	447	0,065	
300	0	0	0,80	0,08	0,12	25	0,003	
010	0,07	0	0,71	0,16	0,06	669	0,098	
210	0,12	0	0,63	0,21	0,03	428	0,062	*BP transition*
20	0,46	0,54	0	0	0	981	0,144	
30	0,08	0	0,68	0,21	0,03	316	0,046	
001	0,59	0	0,32	0,05	0,04	447	0,065	
201	0,76	0	0,17	0,05	0,02	534	0,078	
21	0,77	0	0,18	0,04	0,01	556	0,081	
2	0,64	0,36	0	0	0	1538	0,226	
3	0,77	0	0,17	0,06	0	408	0,06	
4	0,01	0	0,69	0,30	0	104	0,015	

probabilities estimated from an Afrânio Coutinho text

Contexts	Probability							
	0	1	2	3	4	n	freq	
0 0 0	0.21	0.79	0	0	0	195	0.016	
1 0 0	0	0	0.72	0.21	0.07	146	0.011	
2 0 0	0.26	0.74	0	0	0	588	0.048	
3 0 0	0	0	0.67	0.24	0.09	33	0.002	
1 0	0.07	0	0.64	0.21	0.08	2076	0.17	*PE transition*
2 0	0.39	0.61	0	0	0	1516	0.124	
3 0	0.04	0	0.65	0.25	0.06	767	0.063	
1	0.75	0	0.17	0.06	0.02	2781	0.228	
2	0.55	0.45	0	0	0	2781	0.228	
3	0.77	0	0.14	0.08	0.01	997	0.081	
4	0	0	0.70	0.30	0	291	0.023	

probabilities estimated from a João Aguiar text

Besides the two typical trees presented above, other trees appear in some authors. They can be summarized as follows.

1. For 80% of EP texts, but only for 30% of BP texts, (1) is a context.
2. For 100% of EP texts, and for 85% of BP texts, (2) is a context.
3. For 90% of EP texts, and for 80% of BP texts, (3) is a context.

The total set of contexts and families of probability transitions we obtained with our data can be obtained at http://www.ime.usp.br /~tycho/ prosody/ vlmc/ data.

5 Discussion

Let's summarize our main results:

1. Non stressed syllables do not discriminate EP and BP.

   In a very consistent way, we find that the sequences (read from the bottom of the tree to the top) 000, 100, 200 and 300 are contexts in both languages. We can take this fact as expressing the stress pattern of words in Portuguese. 100, for instance, corresponds to a word internal stressed syllable followed by two non stressed syllables. In this case, the next symbol is 2 or 3, i-e the beginning of another word. In contrast, the only possible transitions given 000 (a sequence of three word internal non stressed syllables) are 0, an other non stressed syllable internal to the word, and 1, a word internal stress. Taken together, the probability distribution of the transitions from 100 and 000 reflects the fact that in Portuguese words, a sequence of more than two non stressed syllables can appear before the main stress, but not after it. As expected, 200 patterns with 000, and 300 patterns with 100. As we mentioned above, since BP and EP have the

same lexicon, this is exactly where we expect the two languages to pattern alike.

2. Prosodic word boundaries do not discriminate EP and BP.

   Another similarity expressed by the two tables concerns the prosodic word boundaries. In the great majority of the Brazilian and Portuguese texts, 2 and 3, as well as 20 and 30 are contexts. This can be interpreted as the reflex of a general principle of Lexical integrity at work in Portuguese. The stress pattern of a given word does not depend on informations about the stress pattern of the preceding word in a sentence.

3. Stressed syllables in non initial word boundary do discriminate EP and BP.

   As can be see in the tables and trees, the difference between Brazilian and Portuguese trees lies in the fact that in the latter, but not in the former, 1 is a context. This means that the presence of a stress, in a word internal position, is sufficient to predict the next symbol.

   The probability distributions estimated from an EP text presented in section 4, have the property that after a stressed syllable, we find a non stressed syllable in 75% of the cases, another word beginning with a non-stressed syllable in 17% of the time, another word beginning with a stressed syllable in 6% of the time, and a period in 2% of the time. In the Brazilian text, the probability distribution of the transitions is sensitive to what precedes the stress. We see that the probability of the transition $1 \rightarrow 0$ changes according to whether 1 is preceded by $(0, 0)$ or by $(2, 0)$. The same can be observed in the transitions from the sequence 10.

   This shows that stressed syllables play a role in EP which is absent from BP. This is coherent with the various analyses of EP rhythm mentioned above. First, many of them, on different grounds, (cf. for instance Duarte *et al* 2001, Kleinhenz 1997, Carvalho (1989)) have claimed that EP can be grouped with stress-timed languages like English or German while BP, like other Romance languages, is a syllable-timed language. It is well-known that stressed syllables play a major role in stressed-timed languages. It is striking that the results obtained here points to the same conclusion in the absence of any phonetic information.

   They also corroborate the analysis proposed by Sândalo *et al.* (to appear) in the framework of Optimality Theory. Sândalo et al. show that the different patterns of secondary stress assignment correspond to two different ranking of constraints in which the main difference concerns *Trochee*, which is very high in EP and very low in BP. Again, we can interpret the difference we found between BP and EP using the VLMC approach on written texts as expressing this same difference. In EP, but not in BP, stresses are aligned with the left boundary of the feet. The fact that the

VLMC procedure is sensitive to this feature shows that it is an adequate tool to retrieve rhythmic patterns.

6 Conclusions

In this paper we show that it is possible to retrieve statistical regularities out from codified written texts. These regularities are clearly correlated to previous conjectures concerning rhythmic differences between BP and EP. This was done without using any a priori restriction on the model. This is good point for Statistics and should stimulate further research using more sophisticated probabilistic models in Phonology.

7 Acknowledgements

This work was partially supported by CNPq grants 301301/79 (AG) and 301086/85-0(CG). It is part of PRONEX/FAPESP's Project *Stochastic behavior, critical phenomena and rhythmic pattern identification in natural languages* (grant number 03/09930-9) and also of CNPq's project *Stochastic modeling of speech* (grant number 475177/2004-5). We thank Miguel Galves for writing the Perl software to codify the syllables which was used to treat the data.

References

[1] Carvalho, J.B., "Phonological conditions on Portuguese clitic placement: on syntactic evidence for stress and rhythmic pattern", *Linguistics*, Vol. 27, pp. 405–435, 1989.

[2] Frota, S. and Vigário, M., "Aspectos de prosódia comparada: ritmo e entoação no PE e no PB", *Actas do XV Encontro da Associação Portuguesa de Linguística*, Coimbra, Vol. 1, pp. 533-55, 2000.

[3] Frota, S. and Vigário, M., "On the correlates of rhythm distinctions: the European/ Brazilian Portuguese case", *Probus*, Vol. 13, pp. 247–275, 2001.

[4] Sândalo, F., Abaurre, M. B., Mandel, A. and Galves, C., "Secondary stress in two varieties of Portuguese and the Sotaq optimality based computer program", Probus, Vol. 18, to appear, 2006.

[5] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition, Vol. 73, pp. 265–292, 1999.

[6] Duarte, D, Galves, A., Lopes, N. and Maronna, R., "The statistical analysis of acoustic correlates of speech rhythm", *Workshop on Rhythmic patterns, parameter setting and language change*, ZiF, University of Bielefeld, 2001. Can be downloaded from http://www.physik.uni-bielefeld.de/complexity/duarte.pdf

[7] Rissanen, J., "A universal data compression system", IEEE Trans. Inform. Theory, Vol. 29, pp. 656–664, 1983.

[8] Bühlmann, P. and Wyner, A. J., "Variable length Markov chains", Ann. Statist., Vol. 27, pp. 480–513, 1999.

[9] "The R Project for Statistical Computing", http://www.r-project.org.

[10] Mächler, M., "The VLMC package", 2005. Can be downloaded from http: // cran.r - project. org / doc / packages /VLMC . pdf.

[11] Mächler, M. and Bühlmann, P., "Variable length Markov chains: methodology, computing, and software", J. Comput. Graph. Statist., Vol. 13, pp. 435–455, 2004.

[12] Bühlmann, P., "Model selection for variable length Markov chains and tuning the context algorithm", Ann. Inst. Statist. Math., Vol. 52, pp. 287–315, 2000.

[13] Vigário, M. The prosodic word in European Portuguese, Mouton de Gruyter, 2003.

[14] Kleinhenz, U. "Domain typology at the phonology-syntax interface" in G. Matos et al. (eds) . *Interfaces in Linguistic Theory*, Lisboa: APL/Colibri, pp.201-220, 1997.