Cadernos de
ESTUDOS LINGUÍSTICOS – (59.3), Campinas, pp. 617-630 - set./dez. 2017

LINGUISTIC COMPOSITIONS HIGHLY VOLATILE IN PORTUGUESE

JESÚS ENRIQUE GARCÍA[1]
RAMIN GHOLIZADEH[2]
VERÓNICA ANDREA GONZÁLEZ-LÓPEZ[3]
(UNICAMP)

ABSTRACT: In this paper we use a distance *d* between sequences of *N*-grams to identify *N*-grams that show a different performance when comparing two sequences of *N*-grams. With this tool, we inspect written texts of European Portuguese dated between 16th century and 19th century. We identify the most voluble *N*-grams throughout the period and we also identify *N*-grams that should be considered when studying the linguistic changes from Classical Portuguese to Modern Portuguese. We find that 2-grams composed by unstressed monosyllables followed by paroxytone words (and viceversa) change markedly, from one text to the next, during the whole period. Stressed monosyllabic words (SMW) reveal discrepancies between written texts of the 16th century when compared with texts from the beginning of the 17th century. 2-grams of (i) SMW followed by paroxytone or oxytone word and (ii) paroxytone dissyllabic word or oxytone word followed by a SMW are some of them.
Keywords: *And phrases.* Bayesian information criterion; Partition Markov models; Proximity between *N*-grams.

1. INTRODUCTION

Investigations in the field of historical linguistics record a body of evidence on the changes occurring in written texts, from Classical to Modern Portuguese, including the period from 16th to 19th century. Among these works, Frota et al. (2012)[1] gains relevance as it illuminates the questions raised in this paper. Frota et al. (2012)[1] shows clear evidence of changes occurred from the 16th century to the 17th century, in relation to the prosody of the language. In this paper, our focus is to identify the most relevant linguistic constructions (*N*-grams) that lead to these changes. That is, by sequentially observing the texts, we want to identify the constructions that lead to relevant changes. One line of research is to treat a written text, or some strategic coding thereof, as a sequence of *N*-grams. The structure of *N*-grams plays a very important role in the inspection and modeling of language profiles, see for instance Manning and Schütze (1999)[8]. In some

[1] State University of Campinas, Brazil - jg@ime.unicamp.br
[2] State University of Campinas, Brazil - 1ramin.gholizadh@gmail.com
[3] State University of Campinas, Brazil - veronica@ime.unicamp.br

of the investigations, the focus has been on discriminating between languages, for example see García and González-López (2016)[5] (under the scope of Partition Markov Models (PMM)), in others, the purpose has been to discriminate between varieties of the same language, for example see Galves et al. (2012)[3] (under the scope of Context Tree Models). Moreover, in order to obtain this discrimination, the investigations cited here have centralized their efforts in designing a stochastic profile representative of the language. In adopting this perspective, it is essential to have models that represent *N*-grams and, at the same time models built as general as possible. In this respect, the PMM introduced in García and González-López (2017)[6] fully fulfills this requirement, since PMMs generalize both Markov chains and Context Trees. The PMMs are based on how the state space organizes its strings. The state space is divided into parts of a partition or units, composed by concatenation of elements of the alphabet (strings), all of identical size, which is the order or the memory of the process. Each unit contains strings *s*' and *s* which share the same conditional probability for each element *a* of the process's alphabet. That is, Prob(*a*|*s*)=Prob(*a*|*s*'). An order equal to *N* allows to identify a partition of the set of *N*-grams, where its parts (members of the partition) reveal stochastic synonyms, because the strings in each part share the same conditional probability. This is the kind of representation assumed by the PMM, about the way a language or a process are linearly organized. The first step for building a PMM is to estimate the partition of the state space, i.e. the set of *N*-grams, and as discussed in García and González-López (2017)[6] this is consistently done using the Bayesian Information Criterion (BIC). And according to the results introduced in García and González-López (2015) [4] it is possible to use the BIC as a rule to decide whether two Markov processes follow the same law. That concept can also be used as a similarity criterion between those processes, see García et al. (2017)[7]. In this paper we consider a distance *d* between Markov processes, inspired in the BIC. The advantage in relation to the concept used in García et al.(2017) [7] is that *d* is a distance in the mathematical sense of the term. The values of *d* will be used in linguistic data consisting of written texts of Portuguese, in order to identify the bigrams ($N = 2$) that produce changes. In the specific case under investigation, prosodic changes are identified by means of *d* which compare the frequency distribution of properties related to prosodic word shapes and word stress patterns. Those changes are a consequence of the integration of stress-timing properties into syllable-timed rhythm. In characterizing the languages profiles various authors describe a set of rhythm related properties, syllable structure, variety and complexity, and the properties of stress. Mean word size is also considered as relevant, see for instance Melher and Nespor (2004)[9]. Frota et al. (2012)[1] shows a summary about how those properties are related with a stress-timing language and a syllable-timing language, and the languages between. Based on these large linguistic groups, the European Portuguese is positioned as a language that has gained and lost rhythmic characteristics, transiting between these large groups, without losing its romanic essence. This has led to inspect written texts of Portuguese in search of significant changes, see Frota et al. (2012)[1]. The present paper rescues this problem with the intention of determining which are the bigrams that are shown as variables in the period. We also give a measure which allows to quantify the discrepancy when it is identified.

2. PRELIMINARIES AND NOTATIONS

Here we introduce the measure that we will use, explaining in broad terms its properties. Let (X_t) be a discrete time (order $N < \infty$) Markov chain on a finite alphabet A. Let us call $S = A^N$ the state space and denote the string $a_m a_{m+1} \ldots a_n$ by a_m^n; where $a_i \in A$, $m \leq i \leq n$. For each $a \in A$ and $s \in S$, $P(a|s) = \text{Prob}(X_t = a | X_{t-N}^{t-1} = s)$. In a given sample x_1^n, coming from the stochastic process, the number of occurrences of s in the sample x_1^n is denoted by $Nn(s)$ and the number of occurrences of s followed by a in the sample x_1^n is denoted by $Nn(s, a)$. In this way $\frac{Nn(s,a)}{Nn(s)}$ is the estimator of $P(a|s)$. We will formulate a distance d that, when evaluated in a given string, allows us to decide how far or near the processes are. For instance, suppose a certain linguistic configuration, say, a paroxytone word followed by an unstressed monosyllable, with this criterion we can check if the texts (or processes) are distinguishable or not in relation to such configuration.

Denition 2.1. Consider two Markov chains $(X_{1,t})$ and $(X_{2,t})$ of order N, with finite alphabet A, state space $S = A^N$ and samples $x_{1,1}^{n_1}$, $x_{2,1}^{n_2}$ respectively, define

$$d_{1,2}(s) = \frac{\alpha}{(|A| - 1)\, \ln(n_1 + n_2)} \sum_{a \in A} \left\{ Nn_1(s, a) \ln \left(\frac{Nn_1(s, a)}{Nn_1(s)} \right) + Nn_2(s, a) \ln \left(\frac{Nn_2(s, a)}{Nn_2(s)} \right) - Nn_1 + n_2(s, a) \ln \left(\frac{Nn_1 + n_2(s, a)}{Nn_1 + n_2(s)} \right) \right\}$$

with $Nn_1 + n_2(s, a) = Nn_1(s, a) + Nn_2(s, a)$, $Nn_1 + n_2(s) = Nn_1(s) + Nn_2(s)$, where Nn_1 and Nn_2 are given as usual, computed from the samples $x_{1,1}^{n_1}$ and $x_{2,1}^{n_2}$ respectively. With α real and positive value.

The most relevant properties of d are listed below:

i. The function $d_{1,2}(s)$ is a distance.

If $(X_{i,t})$, $i = 1, 2, 3$ are Markov chains under the assumptions of definition 2.1, with samples $x_{i,1}^{ni}$, $i = 1, 2, 3$ respectively,

$$d_{1,2}(s) \geq 0 \text{ with equality } \Leftrightarrow \frac{Nn_1(s, a)}{Nn_1(s)} = \frac{Nn_2(s, a)}{Nn_2(s)} \quad \forall a \in A,$$

$$d_{1,2}(s) = d_{2,1}(s),$$

$$d_{1,2}(s) \leq d_{1,3}(s) + d_{3,2}(s).$$

ii. Local behavior of process'*s* law.

a. If the stochastic laws of $(X_{i,t})$, $i = 1,2$ in s are the same then $d_{1,2}(s) \underset{\min(n_1,n_2)\to\infty}{\to} 0$. Otherwise $d_{1,2}(s) \underset{\min(n_1,n_2)\to\infty}{\to} \infty$

b. When $\alpha = 2$ and $d_{1,2}(s) < 1$ the stochastic laws of $(X_{i,t})$, $i = 1,2$ are the same, otherwise there are discrepancies.

The usual α value is equal to 2, as described in García and González-López (2017) [6] and introduced in Schwarz (1978) [10]. In this paper we adopt $\alpha = 2$. In order to detect the extreme value of d in S we can define

$$dmax = \max\{d_{1,2}(s), s \in S\}$$

and

$$smax = \arg\max\{dmax\}.$$

If $dmax > 1$, $smax$ is exactly the string we want to recognize, as being relevant in terms of extreme discrepancy, but all the strings with a value $d > 1$ will reveal discrepancies between the processes.

3. DATA

Tycho Brahe corpus is an annotated historical corpus, freely accessible at Galves and Faria (2010)[2]. This corpus uses the chronological criterion of the author's birthdate to assign a time for written texts. The subset of historical written texts included in this study, listed in table 1 is composed by 19 texts from 15 authors, coming from five genres.

Table 1: The set of the Tycho Brahe corpus.

Author	Gândavo	Pinto	Sousa	Brandão	Vieira
Date	1502	1510	1556	1584	1608
Type	narrative	narrative	narrative	narrative	dissertation
Author	Vieira	Vieira	Chagas	Bernardes	Oliveira
Date	1608	1608	1631	1644	1702
Type	letters	sermons	letters	narrative	letters
Author	Aires	Costa	Alorna	Garrett	Garrett
Date	1705	1714	1750	1799	1799
Type	dissertation	letters	letters	letters	narrative
Author	Garrett	Fronteira	Camilo	Ortigão	
Date	1799	1802	1826	1836	
Type	theater	narrative	narrative	letters	

There are previous studies (see Frota et al. (2012)[1]) that show that historical texts such as the listed in table 1 reveal changes in the proportion of occurrence of the placement of the stress in the last or in the penultimate syllable of the word. Also the written texts reveal alterations in the use of monosyllables. These changes are found predominantly from the 16th century to the 17th century. For this reason we guide our inspection to the position in the word occupied by the stress and the word size. Each written text was processed with a slightly modified version of the perl-code "silaba" (by Miguel Galves) that can be freely downloaded for academic purposes at www.ime.usp.br/621tycho/prosody/vlmc/tools/sil4.pl . The software was used to extract two components of each orthographic word, denoted by (i, j), where i is the total number of syllables which integrate the word, $i = 1, 2, ..., 8$ and j indicates the position of the stressed syllable in the word (from left to right). $j = 0$ means no stress in the word. The period (final of sentence) was codified as (0, 0). The alphabet A used here was defined as exposed in table 2. In this approach we used linguistic composition of two words (bigrams), for technical reasons: size of the alphabet and size of the available texts. For example, the linguistic structure 2-7 represents an unstressed monosyllable followed by a *paroxytone* word. The perspective introduced in this study aims to incorporate in the analysis of written texts the dependence between the words that compose them. When considering a bigram s we see that the discrepancies between two written texts will be confirmed, if the next word a to be found in the text 1 and 2, are different. Precisely, given a bigram s, if $d_{1,2}(s) > 1$ we will have that $Prob_1(a|s) \neq Prob_2(a|s)$ with $Prob_i$ computed from the text i, $i = 1, 2$ and a a word of the alphabet.

Table 2: Definition and meaning of each element $a \in A$.

Orthographic word code	a	Meaning
(0,0)	0	final of sentence
(1,1)	1	monosyllable with stress
(1,0)	2	monosyllable without stress
(2,2)	3	dissyllable - stress in the last syllable
(2,1)	4	dissyllable - stress in the first syllable
(i, i), $i \geq 3$	6	*oxytone* word
$(i, i-1)$, $i \geq 3$	7	*paroxytone* word
$(i, i-2)$, $i \geq 3$	8	*proparoxytone* word

4. THE MOST VARIABLE CONFIGURATIONS

Figure 1 shows the values of *dmax* over the years, recorded on the horizontal axis. We note that over the years, the discrepancies detected by *dmax* are more pronounced, except at the end of the 19th century (right). It is worth emphasizing that each discrepancy can be produced by bigrams that are not necessarily identical for all the texts, as detailed in the last column of table 3. We see that the most frequent bigram which produces a high *dmax* (> 1) is

2-4. From the meaning of *d* this means that, the probability conditioned to *an unstressed monosyllable followed by a disyllable with stress at the beginning of the word* motivates such discrepancies between consecutive texts, i.e. these conditional probabilities are very different from a text to the following text, in the cases: 1502-1510, 1584-1608c, 1608c-1631,1608d-1631, 1608s-1631, 1631-1644, 1705-1714, 1714-1750, 1750-1799t, 1750-1799n.

Table 3: Values of *dmax* between a written text and the written text, dated immediately after the previous one.

Date text 1	Date text 2	*dmax*	*smax*
1502	1510	1.08679	2-4
1510	1556	1.64926	2-7
1556	1584	0.71331	1-6
1584	1608c	1.73511	2-4
1584	1608d	1.00874	7-2
1584	1608s	1.35197	2-7
1608c	1631	4.44799	2-4
1608d	1631	3.10919	2-4
1608s	1631	1.78843	2-4
1631	1644	2.00082	2-4
1644	1702	1.03420	4-4
1702	1705	2.04039	7-0
1705	1714	2.59383	2-4
1714	1750	4.46181	2-4
1750	1799c	1.45198	4-7
1750	1799t	6.14052	2-4
1750	1799n	2.29650	2-4
1799c	1802	6.62204	7-2
1799t	1802	6.45512	2-7
1799n	1802	9.04413	7-2
1802	1826	6.39264	7-2
1826	1836	0.52470	2-4

Although this same bigram causes the maximum value of *dmax,* between the texts of 1826 and 1836, the value of *dmax* in this case does not indicate a discrepancy between them (because *dmax* < 1). The bigram 7-2 appears as the next string responsible for discrepancy between texts. The transition probability of *a paroxytone word followed by an unstressed monosyllable* can be considered different for the cases: 1584-1608d, 1799c-1802, 1799n-1802, 1802-1826. Similarly occurs with 2-7, the transition probability of *an unstressed monosyllable followed by a paroxytone word*, can be considered different for the cases: 1510-1556, 1584-1608s, 1799t-1802. In tables 4,5 and 6 we expose the conditional probabilities from the bigram *smax*, to each value of the alphabet *A* and for each pair of texts.

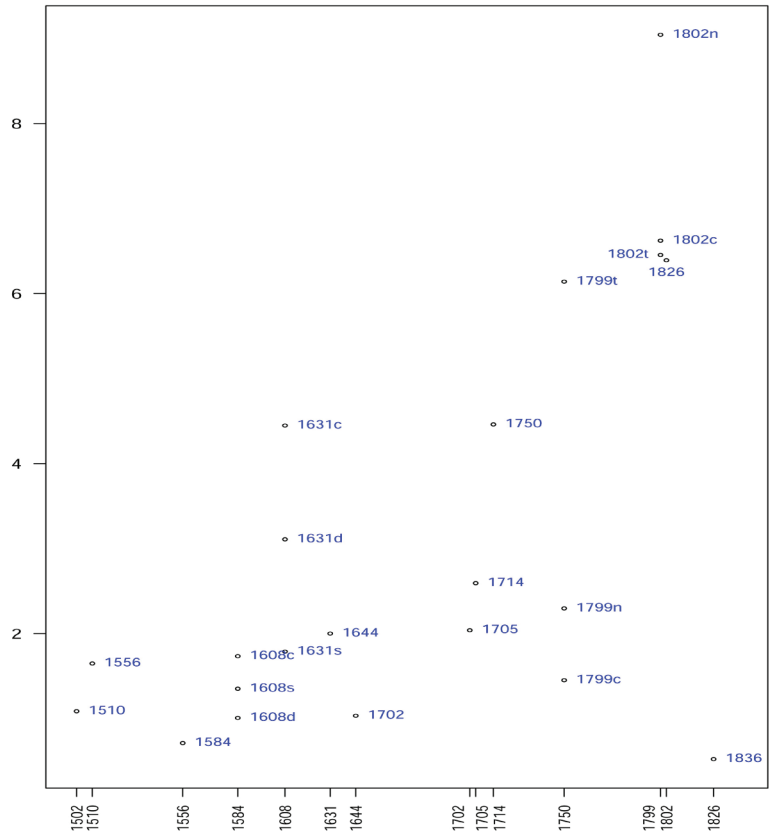

Figure 1: *dmax* values (on the vertical axis) denoted by the year of the second written text (column 2 of the table 3). In the case of a year with several texts, a symbol was attached to the year, which indicates the type of written text: narrative (n), letters (c), sermons (s), theater (t), dissertation (d).

We found cases in which the disparity between the processes is evident, since the conditional probabilities are markedly different, see for instance: 1702-1705, 1714-1750, 1750-1799t, 1750-1799n, 1799c-1802, 1799t-1802, 1799n-1802, 1802-1826. In table 7 we list all the bigrams that show values of $d > 1$ for the cases: 1750-1799t, 1799n-1802 and 1799t-1802 that are those that show a higher *dmax* for the 3 most frequent *smax*, 2-4, 2-7, 7-2. All these cases are in the 18th and early 19th century. We can note that the bigrams (i) *an unstressed monosyllable followed by a disyllable with stress at the beginning of the word*, code: 2-4; (ii) *a paroxytone word followed by an unstressed monosyllable*, code: 7-2 and (iii) *an unstressed monosyllable followed by a paroxytone word*, code: 2-7 detect values of *d* greater than 1, practically in all the written texts, so they should not necessarily be considered as responsible for the changes from the 16th century to the 17th century. We can argue that those constructions between others (also with $d > 1$) are constructions with a tendency to report the particularity of each text.

 doi/10.20396/cel.v59i3.8651002

Table 4: Conditional probabilities from *smax* to each element *a* of the alphabet *A*.
Texts: 1502, 1510, 1556, 1584, 1608c, 1608d, 1608s, 1631.

a	*smax*: 2-4 1502	(1.08679) 1510	*smax*: 2-7 1510	(1.64926) 1556
0	0.03251	0.01590	0.01974	0.05232
1	0.06572	0.04623	0.04877	0.04559
2	0.34383	0.39770	0.52833	0.52970
3	0.04220	0.04310	0.05864	0.03270
4	0.28087	0.23577	0.19509	0.16385
6	0.02421	0.01485	0.01635	0.01944
7	0.19301	0.23159	0.12602	0.13896
8	0.01764	0.01485	0.00705	0.01744

a	*smax:* 2-4 1584	(1.73511) 1608c	*smax:* 7-2 1584	(1.00874) 1608d	*smax:* 2-7 1584	(1.35197) 1608s
0	0.03356	0.02277	0.00022	0.00000	0.05627	0.11444
1	0.07102	0.06624	0.10044	0.09927	0.05447	0.05488
2	0.38474	0.31714	0.13624	0.15025	0.52607	0.47124
3	0.04528	0.04251	0.07707	0.04350	0.03614	0.03284
4	0.20363	0.20402	0.32729	0.29417	0.15643	0.16809
6	0.02528	0.02638	0.03930	0.03986	0.01859	0.02203
7	0.21719	0.31144	0.29127	0.34266	0.13552	0.12464
8	0.01931	0.00949	0.02817	0.03028	0.01652	0.01183

a	*smax:* 2-4 1608c	(4.44799) 1631	*smax:* 2-4 1608d	(3.10919) 1631	*smax:* 2-4 1608s	(1.78843) 1631
0	0.02277	0.04852	0.03027	0.04852	0.06825	0.04852
1	0.06624	0.08622	0.07731	0.08622	0.07620	0.08622
2	0.31714	0.32035	0.33885	0.32035	0.35255	0.32035
3	0.04251	0.09578	0.03550	0.09578	0.04473	0.09578
4	0.20402	0.23178	0.21733	0.23178	0.22949	0.23178
6	0.02638	0.01768	0.03223	0.01768	0.02175	0.01768
7	0.31144	0.19012	0.24913	0.19012	0.19307	0.19012
8	0.00949	0.00956	0.01938	0.00956	0.01397	0.00956

Table 5: Conditional probabilities from *smax* to each element *a* of the alphabet *A*.
Texts: 1631, 1644, 1702, 1705, 1714, 1750, 1799c, 1799t, 1799n.

a	*smax:* 2-4 1631	(2.00082) 1644	*smax:* 4-4 1644	(1.03420) 1702	*smax:* 7-0 1702	(2.04039) 1705
0	0.04852	0.03472	0.03494	0.11220	0.00000	0.00000
1	0.08622	0.07377	0.10519	0.09368	0.06766	0.12903
2	0.32035	0.33500	0.31225	0.30174	0.53498	0.25605
3	0.09578	0.04860	0.05112	0.07190	0.05791	0.05645
4	0.23178	0.23085	0.22030	0.18954	0.22764	0.30847
6	0.01768	0.02864	0.03715	0.02832	0.00516	0.03427
7	0.19012	0.22282	0.21552	0.18736	0.09117	0.15323
8	0.00956	0.02560	0.02354	0.01525	0.01548	0.06250

a	*smax*: 2-4 1705	(2.59383) 1714	*smax*: 2-4 1714	(4.46181) 1750
0	0.07313	0.04440	0.04440	0.03273
1	0.09521	0.07985	0.07985	0.05687
2	0.33216	0.33246	0.33246	0.29573
3	0.03062	0.10373	0.10373	0.03412
4	0.23695	0.23657	0.23657	0.19684
6	0.01824	0.01493	0.01493	0.02484
7	0.20030	0.17761	0.17761	0.34123
8	0.01339	0.01045	0.01045	0.01764

a	*smax:* 4-7 1750	(1.45198) 1799c	*smax:* 2-4 1750	(6.14052) 1799t	*smax:* 2-4 1750	(2.29650) 1799n
0	0.08432	0.10345	0.03273	0.13364	0.03273	0.06361
1	0.05295	0.05314	0.05687	0.10649	0.05687	0.07905
2	0.40000	0.44432	0.29573	0.27959	0.29573	0.30952
3	0.12310	0.03957	0.03412	0.04200	0.03412	0.05047
4	0.17774	0.16789	0.19684	0.24608	0.19684	0.21318
6	0.02226	0.02600	0.02484	0.01273	0.02484	0.02374
7	0.11872	0.14811	0.34123	0.17522	0.34123	0.23070
8	0.02091	0.01752	0.01764	0.00424	0.01764	0.02973

Table 6: Conditional probabilities from *smax* to each element *a* of the alphabet *A*. Texts: 1799c, 1799t, 1799n, 1802, 1826.

a	*smax:* 7-2 1799c	(6.62204) 1802	*smax:* 2-7 1799t	(6.45512) 1802	*smax:* 7-2 1799n	(9.04413) 1802
0	0.00064	0.00031	0.25458	0.06325	0.00183	0.00031
1	0.12162	0.06796	0.08215	0.03171	0.10353	0.06796
2	0.16651	0.38500	0.35777	0.51374	0.15735	0.38500
3	0.05890	0.04704	0.04684	0.04228	0.05153	0.04704
4	0.28335	0.22984	0.14053	0.14711	0.29684	0.22984
6	0.04011	0.02673	0.01154	0.02872	0.02863	0.02673
7	0.31137	0.22770	0.10251	0.16138	0.32410	0.22770
8	0.01751	0.01542	0.00407	0.01180	0.03619	0.01542

a	*smax*: 7 -2 1802	(6.39264) 1826
0	0.00031	0.00000
1	0.06796	0.08152
2	0.38500	0.13136
3	0.04704	0.05919
4	0.22984	0.32347
6	0.02673	0.03998
7	0.22770	0.33022
8	0.01543	0.03427

Table 7: Cases with bigger values of *d* and different *smax*: 1750-1799t, 1799n-1802, 1799t-1802. In bold the bigrams that most often produce the highest values of *d*.

$d_{1750,\ 1799t}(s)$	*s*	$d_{1799n,\ 1802}(s)$	*s*	$d_{1799t,\ 1802}(s)$	*s*
1.12915	2-2	1.00300	4-7	1.05268	2-6
1.13272	7-4	1.21533	1-2	1.13008	4-3
1.18684	4-1	1.39296	6-2	1.16699	4-1
1.21775	1-4	1.55913	2-2	1.23502	7-0
1.22487	3-2	**1.69919**	**2-7**	1.31616	1-2
1.22692	1-7	1.96170	3-2	1.51818	0-2
1.26838	6-2	**2.01062**	**2-4**	1.62364	1-4
1.29315	7-7	2.08092	4-4	1.62953	7-7
1.39431	1-2	5.31407	4-2	**1.66957**	**7-2**
2.18292	2-1	**9.04413**	**7-2**	1.75069	1-7
2.35651	4-4			1.77862	7-4
3.37187	4-2			2.56165	2-1

3.38483	**7-2**			2.58853	2-2
4.00310	4-7			3.62165	4-2
4.66029	**2-7**			4.05870	4-4
6.14052	2-4			4.39322	4-7
				6.39768	**2-4**
				6.45512	**2-7**

5. FROM THE 16TH CENTURY TO THE BEGINNING OF THE 17TH CENTURY

In Frota et al.(2012)[1] significant changes are reported in the language from the 16th century to the 17th century. In the previous section we noticed that some bigrams are intrinsically variable, being characterized by large values of *dmax*. In this section, we examine the transition from the 16th century to the 17th century, taking into account that the 3 configurations cited in the previous section do not necessarily lead to drastic changes in the language. We record all the bigrams which report changes (i.e. with values of $d > 1$), considering each written text of the 16th century in relation to the written texts dated immediately afterwards, until the beginning of the 17th century. Tables 8, 9 and 10 present the results. In table 11 we list the bigrams detected as a change in the comparison between each written text of the 16th century with the first 3 written texts of the 17th century: 1608c, 1608d and 1608s. In that list we exclude configurations that are identified as changes between texts of the 16th century itself.

Table 8: Values of d and bigrams such that $d > 1$, between texts of the 16th century and Vieira's texts: 1608c, 1608d and 1608s. In bold letter the most frequent bigrams, according to the previous section.

$d_{1502,\ 1510}(s)$	s	$d_{1502,\ 1556}(s)$	s	$d_{1502,\ 1584}(s)$	s
1.08679	**2-4**	1.10873	**7-2**	1.10897	4-2
		1.51328	4-2	1.32348	4-4

$d_{1502,\ 1608c}(s)$	s	$d_{1502,\ 1608d}(s)$	s	$d_{1502,\ 1608s}(s)$	s
1.15103	4-2	1.85333	4-2	1.07436	1-4 (V)
1.21610	2-3 (II)	1.93021	**7-2**		
2.16232	**2-4**				

Table 9: Values of d and bigrams such that $d > 1$, between texts of the 16th century and Vieira's texts: 1608c, 1608d and 1608s. In bold letter the most frequent bigrams, according to the previous section.

$d_{1510,\ 1556}(s)$	s	$d_{1510,\ 1584}(s)$	s
1.39871	4-2	1.02257	4-0
1.64242	**7-2**	1.04342	**2-4**
1.64926	**2-7**	1.06105	1-3
		1.12283	**7-2**
		1.15678	4-4
		1.19666	3-6
		1.33785	1-6
		1.37050	7-0
		1.51808	**2-7**

$d_{1510,\ 1608c}(s)$	s	$d_{1510,\ 1608d}(s)$	s	$d_{1510,\ 1608s}(s)$	s
1.20665	**2-7**	1.04371	1-7 (VI)	1.41470	4-4
2.03579	**2-4**	1.07434	1-4 (V)	2.06415	4-7 (I)
		1.13177	7-0	3.20374	**2-4**
		1.31517	1-6	4.43740	**2-7**
		1.36703	4-2		
		1.55650	**2-4**		
		1.81723	**2-7**		
		2.43819	**7-2**		

Table 10: Values of d and bigrams such that $d > 1$, between texts of the 16th century and Vieira's texts: 1608c, 1608d and 1608s. In bold letter the most frequent bigrams, according to the previous section.

$d_{1556,\ 1608c}(s)$	s	$d_{1556,\ 1608s}(s)$	s
1.32334	**2-4**	1.57438	4-7 (I)
1.66714	**7-2**	1.67042	**2-4**
		1.94014	**2-7**

$d_{1584,\ 1608c}(s)$	s	$d_{1584,\ 1608d}(s)$	s	$d_{1584,\ 1608s}(s)$	s
1.07605	3-1 (III)	1.00874	**7-2**	1.07634	**2-4**
1.10511	6-1 (IV)			1.14234	3-1 (III)
1.10537	**7-2**			1.35197	**2-7**
1.14962	1-6				
1.23217	2-3 (II)				
1.41250	3-6				
1.73511	**2-4**				

Table 11: Bigrams that announce changes between texts of the 16th century when compared to texts of beginning of the 17th century: 1608c, 1608d, 1608s. In the third column are indicated the cases covered by the configuration, see tables 8, 9, 10.

Code string	Bigram	Reference
4-7	a disyllable with stress on the first syllable followed by a *paroxytone* word	(I)
2-3	an unstressed monosyllable followed by a disyllable with stress on the last syllable	(II)
3-1	a disyllable with stress on the last syllable followed by a stressed monosyllabic word	(III)
6-1	an *oxytone* word followed by a stressed monosyllabic word	(IV)
1-4	a stressed monosyllabic word followed by a disyllable with stress on the first syllable	(V)
1-7	a stressed monosyllabic word followed by a *paroxytone* word	(VI)

6. CONCLUSIONS

In this work we introduce a strategy to identify linguistic structures (bigrams) that generate alterations of the Portuguese. Also it is possible to identify the bigrams more strongly associated with historical changes. Bigrams with large values of *d* unrelated to temporal changes could possibly be used to discriminate linguistic genres or particular aspects of texts. Moreover, the idea of identifying the language with sequences of *N*-grams thus adopting the measure *d* to proceed to the detection of changes, can be applied to other contexts and problems, helping to solve and review linguistic alterations proclaimed in the literature of the area of historical linguistics. In this instance, it is necessary to make some observations. The *dmax* detects volatile linguistic constructions that expose changes in several moments from Classical Portuguese to Modern Portuguese (period: 16th century to 19th century). Among them, the most outstanding constructions, with maximum *d* value and most frequent, are: (i) *an unstressed monosyllable followed by a paroxytone disyllable word*, (ii) *a paroxytone word followed by an unstressed monosyllable* and (iii) *an unstressed monosyllable followed by a paroxytone word*. These voluble linguistic constructions allow to delineate the profile of the Portuguese language in the period: 16th century to 19th century, showing in a clear way the constructions more associated to the changes of the period. These results already show that bigrams composed by unstressed monosyllables and paroxytone words (and viceversa) are the most likely to suffer alteration. It should be remembered that in Frota et al.(2012)[1] these two characteristics indicate significant changes in the Portuguese of the period: 16th-17th. In the present work we go further, because bigrams take into account the dependence between both aspects: unstressed monosyllables and paroxytone words. When comparing the texts of the 16th century with the first texts

of the 17th century, it is possible to detect a series of bigrams that indicate important differences, since, in these cases, the measure *d* adopts values greater than 1. These are (a) *a disyllable with stress on the first syllable followed by a paroxytone word* (indicated by two texts), (b) *an unstressed monosyllable followed by a disyllable with stress on the last syllable* (indicated by two text), (c) *a disyllable with stress on the last syllable followed by a stressed monosyllabic word* (indicated by two texts), (d) *an oxytone word followed by a stressed monosyllabic word* (indicated by one text), (e) *a stressed monosyllabic word followed by a disyllable with stress on the first syllable* (indicated by two texts) and (f) *a stressed monosyllabic word followed by a paroxytone word* (indicated by one text). This type of study could motivate others that allow in fact to identify precisely how the prosody concerns intonational and rhythmic patterns involving stress alternation in a language.

REFERENCES

[1] S. Frota, C. Galves, M. Vigario, V. A. González-López and B. Abaurre, The phonology of rhythm from Classical to Modern Portuguese, Journal of Historical Linguistics (2012) 2.2 173-207.

[2] C. Galves and P. Faria, Tycho Brahe Parsed Corpus of Historical Portuguese. http://www.tycho.iel.unicamp.br/ tycho/corpus/en/index.html (2010).

[3] A. Galves, C. Galves, J. García, N. L. Garcia and F. Leonardi, Context tree selection and linguistic rhythm retrieval from written texts, The Annals of Applied Statistics (2012) 6(1) 186-209.

[4] Jesus E. García and V. A. González-López, Detecting regime changes in Markov models, New Trends in Stochastic Modeling and Data Analysis (2015) (in chapter 2, p. 103).

[5] Jesus E. García and V. A. González-López, Optimal Partition of Markov Models and Automatic Classification of Languages, Stochastic and Data Analysis Methods and Applications in Statistics and Demography (2016) (in chapter 5, p. 207).

[6] Jesus E. García and V. A. González-López, Consistent Estimation of Partition Markov Models, Entropy (2017) 19(4) 160.

[7] Jesus E. García, V. A. González-López and F. H. Kubo de Andrade, Dissimilarity between Markovian Processes Applied to Industrial Processes, AIP Conference Proceedings (2017)1863 220002.

[8] C.D. Manning and H. Schütze, Foundations of statistical natural language processing, Vol. 999. Cambridge: MIT press, (1999).

[9] J. Mehler and M. Nespor, Linguistic rhythm and the acquisition of language, Vol. 3, pp. 213-222. Oxford: Oxford University Press, (2004).

[10] G. Schwarz, Estimating the dimension of a model, The annals of statistics, (1978) 6(2) 461-464.

Jesus E. García: Department of Statistics, University of Campinas, Campinas, SP, CEP 13083-859, Brazil - *E-mail address*: jg@ime.unicamp.br

R. Gholizadeh: University of Campinas, Campinas, SP, CEP: 13083-859, Brazil - *E-mail address*: 1ramin.gholizadh@gmail.com

V. A. González-López: Department of Statistics, University of Campinas, Campinas, SP, CEP: 13083-859, Brazil - *E-mail address*: veronica@ime.unicamp.br