What Rhythmic Signature Says About Poetic Corpora

Adiel Mittmann
Universidade Federal de Santa Catarina
Florianópolis, Brazil
adiel@mittmann.net.br

Paulo Henrique Pergher
Universidade Federal de Santa Catarina
Florianópolis, Brazil
paulopergher@gmail.com

Alckmar Luiz dos Santos
Universidade Federal de Santa Catarina
Florianópolis, Brazil
alckmar@gmail.com

Abstract

The distribution of stressed syllables within a verse are fairly varied in Portuguese and Spanish, which allows poets to use rhythmic patterns in specific ways. Longer verses do tend to impose constraints on which syllables are allowed to carry stress, but there is still room for variation; for instance, verses with 10 syllables can belong to one of 22 standard patterns. In this article, we analyze various sets of verses in order to establish their relative frequency distribution of rhythmic patterns, which we term the rhythmic signature, and to understand how such signatures relate to each other. We also provide recommendations for using two different methods of comparing them: the Gini coefficient and dendrograms. Three experiments are conducted on a corpus with more than 250,000 verses: first, we compare poets who wrote in Portuguese; then, we study several books written by a single author; finally, we compare poets from two languages, Portuguese and Spanish. Our results show that many authors have a distinguishable rhythmic signature, to the point of being fairly unique. It is also clear from our results that, with deliberate intention, a poet can mimic the rhythmic style of another. Finally, the third experiment indicates that verses in Portuguese and Spanish are rhythmically similar and that differences in style can mostly be ascribed to the authors. Overall, our results show that the rhythmic signature aids in characterizing the style of verses, which gives experts objective, reliable information from which deeper insights can be obtained.

1 Introduction

In Portuguese, Spanish and other languages, the placement of stressed syllables within a verse can vary widely, and most of the variation is a choice, conscious or unconscious, made by the poet. Longer verses tend to impose certain restrictions, but usually there is ample room left for the poet to choose different rhythmic patterns of strong and weak syllables.

Automatic scansion tools are able to scan a large body of verses quickly, which allows us to analyze from a distance (Moretti 2013) the different rhythmic patterns used by poets. Instead of looking at the rhythmic pattern of a few verses, we can instead look at the broad strokes painted by poems, poets, languages, periods of time: which patterns are favored, with which variety they are used, how different sets of verses can be hierarchically grouped, and so on. In the past, such efforts were very limited in their scope due to the prohibitive cost of manually scanning thousands of verses, though they do exist (Chociay 1994).

In this article, we explore the relative frequency distribution of rhythmic patterns, which we call the rhythmic signature. We define it, establish some ways to deal with the uncertainty inherent to such distributions and provide recommendations on using two ways of comparing signatures: the Gini coefficient and dendrograms. In order to provide an answer to this article's title, we use rhythmic signatures to examine three different scenarios: verses written by many different poets in one language, Portuguese; verses written by a single poet over a stretch of 54 years; and verses written in Portuguese and Spanish by several poets.

The rhythmic signatures used in this article, as well as many tables and figures, were computed by Aoidos (Mittmann–von Wangenheim–Luiz dos Santos 2016), our automatic scansion system capable of scanning verses in Portuguese and Spanish. Aoidos takes as input files in the TEI XML format and produces several analyses, from scansion itself to tables of rhythmic signatures and metaplasm usage. It is available online at https://aoidos.ufsc.br/. We have already used Aoidos in order to compare automatic results to those published by experts in the field of versification (Mittmann–Maia 2017) and to analyze epic verse in Portuguese (Mittmann–Luiz dos Santos 2018), when we used for the first time rhythmic signatures. Similar systems exist for Spanish (Navarro-Colorado 2016) and other close languages, such as French (Beaudouin–Yvon 1996; Delente–Renault 2015).

The remainder of this article is structured as follows. Sect. 2 provides information on the corpus used both for the experiments and for the analyses carried out in the next section; Sect. 3 defines the rhythmic signature and its properties; Sect. 4 presents three different experiments based on the rhythmic signature; in Sect. 5, we discuss some insights that emerged from the experiments; and finally, Sect. 6 contains our concluding remarks.

2 Corpus

The corpus used in this article is composed of 286,388 verses, as summarized in TAB. 1. The majority of verses (76.9%) are 10 syllables long, with a sizable minority (15.5%) of verses with 7 syllables. Only these two meters are considered in this article; the remaining verses (7.6%) are not used. Almost all (99.91%) of Spanish verses are 10-syllable long; the remaining ones (0.09%) are mostly 6-syllable verses that introduce *estrambotes*.

In this article, we measure the number of syllables in a verse according to the scheme currently prevalent in Portuguese versification studies—even when we are considering verses in Spanish. Syllables in a verse are thus only counted up until the last stressed one, so that a verse whose last stressed syllable is the 10th (hence in Portuguese they are called *decassílabos*) is said to have 10 syllables, even though it most often contains a further unstressed syllable (hence in Spanish they are called *endecasílabos*).

TAB. 2 presents details on the verses written in Portuguese. There are 116,912 verses with 10 syllables and 44,400 verses with 7 syllables. Of those with 10 syllables, 52.8% belong to ten epic poems, which are long and typically written in this meter. Of those with 7 syllables, 64.6% were written by only one poet, Gregório de Matos—this high proportion should not taint our analyses, as long as we take the necessary precautions. Poet Bastos Tigre, who wrote 16.7% of the verses in Portuguese, originally had an additional 6,106 verses in the corpus, which were subsequently removed from all analyses due to being repeated and are not counted anywhere in this article. The complete works of Bastos Tigre does not include another copy of *Bromilíadas*, which appears separately in the table; furthermore, there is at least one small book that is not included here. One item in the corpus was written by two authors: that is the book *Pimentões*, written by Olavo Bilac and Guimarães Passos; the rhythmic signature derived from this book must be interpreted as an amalgam to which two authors contributed. There is one translation in the corpus: that is Dante's *Divine Comedy*, as translated by Xavier Pinheiro.

TAB. 3 summarizes the verses in Spanish. There are two epic poems, *La Araucana* by Alonso de Ercilla and *La Argentina* by Martín del Barco Centenera, which together make up 44.7% of the corpus. The remaining verses are sonnets from the Spanish Golden Age, which were collected and published in XML form by Navarro-Colorado–Lafoz–Sánchez (2016); this collection of sonnets is referred to as the Siglo de Oro corpus in our article. Most sonnets in the Siglo de Oro corpus are written in Spanish, but there are two written in Portuguese, by Quevedo and Borja, which were included in TAB. 2 for the sake of completion. Although poet Gregório de Matos wrote most of his poems in Portuguese, some of them he wrote in Spanish; our own versions of his poems in Spanish are not included in the corpus because the archaic spelling has not been updated yet, but the Siglo de Oro corpus does include at least some of his sonnets written in Spanish.

Verses with a length other than 7 or 10 are included in TAB. 2 and TAB. 3, though they are not used in our analyses, for two reasons. First, so that the general picture of

	7	10	Other	Total
Portuguese	44,400	116,912	21,692	183,004
Spanish	4	102,825	92	102,921
Total	44,404	219,737	21,784	285,925

TAB. 1: Number of verses in the corpus, by language and meter

Poet(s)	Birth	Works	7	10	Other	Total
L. de Camões	1524	**(L)** Epic: *Os Lusíadas*		8,816		8,816
F. de Quevedo	1580	Sonnet		13	1	14
F. de Borja	1581	Sonnet		14		14
Sá de Meneses	1600	**(M)** Epic: *Malaca Conquistada*		10,635	21	10,656
G. de Matos	1636	Complete works	28,692	4,070	1,385	34,147
S. R. Durão	1722	**(C)** Epic: *Caramuru*		6,672		6,672
C. M. da Costa	1729	**(V)** Epic: *Vila Rica*		2,717	1	2,718
		Misc. Book: *Obras Poéticas*	231	5,736	1,715	7,682
B. da Gama	1741	**(U)** Epic: *O Uraguai*		1,377		1,377
T. A. Gonzaga	1744	Satirical: *Cartas Chilenas*		4,172	12	4,184
G. de Magalhães	1811	Misc. Book: *Suspiros Poéticos*	540	3,790	1,307	5,637
X. Pinheiro	1822	Narrative: *Divina Comédia*		14,226	7	14,233
G. Dias	1823	**(T)** Epic: *Os Timbiras*		2,004	28	2,032
F. Varela	1841	**(A)** Epic: *Anchieta*		8,480	22	8,502
D. Silveira	1854	Complete works	4,082	4,490	3,139	11,711
A. Figueredo	1864	Several books	1,022	2,760	4,451	8,233
Bilac, Passos	1865	Satire: *Pimentões*	1,502			1,502
B. Tigre	1882	**(R)** Epic: *Bromilíadas*		3,308	2	3,310
		Complete works	8,055	10,014	9,117	27,186
A. dos Anjos	1884	Complete works	276	5,842	476	6,594
C. A. Nunes	1897	**(B)** Epic: *Os Brasileidas*		8,503	1	8,504
J. Teixeira	19??	**(F)** Epic: *Famagusta*		9,273	7	9,280
Total			44,400	116,912	21,692	183,004

TAB. 2: Number of verses in Portuguese, by poet and meter

Poet/Group	Birth	Works	7	10	Other	Total
A. de Ercilla	1533	Epic: *La Araucana*		21,072		21,072
M. de Centenera	1535	Epic: *La Argentina*		10,727	17	10,744
Siglo de Oro	1398-1672	Sonnets	4	71,026	75	71,105
Total			4	102,825	92	102,921

TAB. 3: Number of verses in Spanish, by poet or group of poets and meter

which meters a work employs is still visible; thus, Gonçalves Dias' *Os Timbiras*, though an epic poem with mostly 10-syllable verses, does contain verses of other types. Second, so that scansion errors due to underlying problems in the text itself do not interfere in the total amount of verses in a work: such is the case with Sá de Meneses' *Malaca Conquistada*, which contains 10,656 verses, even if 21 could not be scanned in 10 syllables.

3 Rhythmic signature

3.1 Definition

The rhythmic signature of a set of same-length verses is the relative frequency distribution of the rhythmic patterns of the verses in the set. The rhythmic pattern of a verse describes the position of stressed syllables within it. For instance, these are the five verses that begin Carlos Alberto Nunes' *Os Brasileidas* and their respective rhythmic patterns:

1-3-6-10	*Musa, canta-me a régia poranduba*
3-6-10	*Das bandeiras, os feitos sublimados*
3-6-8-10	*Dos heróis que o Brasil plasmar souberam*
2-6-10	*Través do Pindorama, demarcando*
3-6-10	*Nos sertões a conquista e as esperanças.*

The rhythmic signature of this set of verses is simply the frequency with which each rhythmic pattern occurs:

1-3-6-10	**3-6-10**	**3-6-8-10**	**2-6-10**
16.7%	33.3%	16.7%	16.7%

The exact manner in which one determines the rhythmic pattern of a verse is not entirely absolute. Some experts would argue that the stress of certain words would become weak before a stronger word, even if the two stressed syllables are not next to each other. Drawing arguments from phonological features of Portuguese, others would claim that a rhythmic pattern like 6-10 is impossible, because a secondary stress will invariably become dominant in one of the first four unstressed syllables. In any case, Aoidos computes rhythmic patterns by following a few simple rules, like "a stressed syllable cannot follow a stressed syllable". Unless any such rule would be broken, a word is always stressed in the same way and no secondary stress is ever taken into account; phenomena like systoles and diastoles are treated in a way that it does not affect the computing of rhythmic patterns.

3.2 Uncertainty

The rhythmic signature can be calculated for sets of any size, but the smaller the set, the less trustworthy the information is. In the example above, the rhythmic signature was calculated for only 5 verses; one can hardly expect that this signature is capable of representing the whole epic poem or even one of its cantos. One way of dealing with this uncertainty is by using statistical confidence intervals, which, in this case, are given by

$$p \pm z \sqrt{\frac{p(1-p)}{n}},$$

where p is the proportion of verses that follow a given rhythmic pattern, n is the total number of verses in the set and z selects the desired confidence level (so that, for instance, for 95% confidence one would have $z = 1.960$). Aoidos, when presenting results to users, currently calculates confidence intervals and then employs a mix of different font sizes and colors to communicate how reliable the figures in a rhythmic signature are.

There are situations, however, when one would like to go beyond looking at the numbers; that is the case when clustering techniques are employed. In order to estimate how much a rhythmic signature can be trusted given the quantity of verses it was based on, an analysis was undertaken: we considered all 10- and 7-syllable verses from books or poems written in Portuguese in our corpus. We then established a set of eight sample sizes, ranging from 10 to 5,000. For each sample size, each book or poem was considered in turn: 100 random samples were taken and the Euclidean distance was calculated between the rhythmic signature of the random sample and the rhythmic signature of the book or poem as a whole. Sample sizes larger than 50% of the whole book or poem were not considered. The whole procedure was then repeated for several minimum file sizes, again ranging from 10 to 5,000 verses. TAB. 4 shows the *maximum* distance between a random sample and the book it was taken from, according to different sample sizes. TAB. 5 does the same, but shows *mean* distances instead.

According to TAB. 4 and TAB. 5, if we consider a set of 1,000 verses, then we can expect that the distance from its rhythmic signature to that of the whole should be below 0.03 on average and not much larger than 0.05 in the worst case. They also alert us to the fact that small sets of verses should be handled with care: given 100 verses, on average a distance of almost 0.10 is to be expected from the rhythmic signature of the sample to that of the whole, and, in the worst case, this distance could reach to almost 0.20.

In practice, we will not be calculating rhythmic signatures from random samples extracted from a larger set; we will always use the entire set that we wish to analyze. In such cases, we can consider that our entire set is indeed a sample size of all the (infinite) verses that the poet could have written when that specific poem was being written.

When using rhythmic signatures, it is also important to consider which specific rhythmic patterns can be usefully analyzed. When one considers smaller verses, typically up to 7 syllables, the internal distribution of stressed syllables is completely up to the poet: there are no restrictions. Larger verses, typically from 8 to 12 syllables, on the other hand, usually coordinate smaller units. TAB. 6 shows the number of theoretical and actual rhythmic patterns for meters from 1 to 12. It is evident that deviations can occur, and poets have been exploiting them for centuries; therefore, the actual number of rhythmic patterns refers only to the most commonly found rhythms. One notices that many sizes allow enough patterns that a meaningful rhythmic signature can be computed and analyzed.

	10	100	250	500	750	1,000	2,500	5,000
7	0.630	0.183	0.108	0.078	0.056	0.048	0.022	
10	0.627	0.180	0.111	0.074	0.066	0.052	0.029	0.017

TAB. 4: Maximum distance between a random sample size and the book or poem it was taken from

	10	100	250	500	750	1,000	2,500	5,000
7	0.286	0.085	0.051	0.034	0.027	0.024	0.014	
10	0.297	0.091	0.057	0.039	0.032	0.027	0.016	0.010

TAB. 5: Mean distance between a random sample size and the book or poem it was taken from

	1	2	3	4	5	6	7	8	9	10	11	12
Theoretical	1	1	2	3	5	8	13	21	34	55	89	144
Actual	1	1	2	3	5	8	13	6	11	22	25	52

TAB. 6: Theoretical and actual number of possible rhythmic patterns, according to meter

Of particular interest to this article is the actual number of rhythmic patterns that 10-syllable verses may adopt: 22, instead of the theoretical 55. In practice, therefore, we can expect to find 22 more common, "allowed" rhythmic patterns, and 33 less important patterns. The latter group will always exist due to problems in the text, scansion errors produced by Aoidos, intentional "breaking of the rules" by poets, etc. In fact, in our corpus, 51 of the 55 theoretical types can be found.

Considering which rhythmic patterns are theoretical and which actually occur commonly in practice is important because certain statistical indices, such as the Gini index introduced in the next section, will be skewed when additional patterns are included in the signature. It must be noted that this is not the case, e.g., when clustering techniques are used, because the distance between signatures remain mostly unaffected (the addition of an extra dimension which has a zero value for all rhythmic signatures changes nothing).

3.3 The Gini coefficient

The Gini coefficient or index was originally proposed in order to measure inequality of income. It can, however, measure the inequality among any values in a frequency distribution—such as a rhythmic signature. One advantage it has over statistical measures such as standard deviation is that it varies from 0 (complete equality) to 1 (complete inequality). Whereas in its usual application the values are the incomes of citizens, in our case the values are the quantity of verses (or their proportion) that belong to each rhythmic pattern. As the discussion in the previous section warns, care must be taken to select which rhythmic patterns can be considered true members of the population.

For our purposes, we can interpret high Gini coefficients to mean that a few rhythmic patterns dominate the poet's writing: there is a strong preference for a small number of rhythms. A low Gini coefficient, on the other hand, indicates that the poet uses a wider variety of rhythms, perhaps employing rarer rhythms with a greater frequency.

The calculation of the Gini index is simple. It is given by

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left| x_i - x_j \right|}{2n \sum_{i=1}^{n} x_i},$$

where n is the number of rhythmic patterns and x_i is the frequency (or absolute number) of verses that adopt that rhythmic pattern.

3.4 Dendrograms

One way of exploring the relationships among different rhythmic signatures is building a dendrogram, a traditional hierarchical clustering technique. When using dendrograms, there is no need to exclude the theoretical rhythmic patterns, since they, by definition, are rare and will not significantly affect the clusterization algorithm. For the sake of simplicity, this section will assume that each rhythmic signature is calculated from the verses of a single poet, so that we can talk about the more concrete "poets", rather than "verse sets".

The dendrogram algorithm takes an important parameter, which is the linkage method. It specifies how the similarity between two groups of poets (which could be groups of just one poet) is calculated. There are three methods of particular interest here:

- Single linkage. In order to compare two groups of poets, this method selects the two poets (one from each group) that look the most similar, and then considers that the two groups as a whole are as similar to each other as those two poets.
- Complete linkage. This method is similar to single linkage, but it takes the two poets (one from each group) that are the most different from each other.
- Average linkage. This method does not select any poets; it instead creates an average poet for each group, made by averaging all rhythmic signatures. The two average poets are then compared.

Each method has its advantages and disadvantages. Single linkage is very intuitive, since it chooses the two most similar poets in order to compare groups; but the end result usually looks like one group of poets, to which all poets are added, one by one, until one big group remains. This means that usually there is no cut that can be made to compare different groups of poets; any one such cut ends up dividing the poets into one big group and several one-poet groups. Complete linkage is not as intuitive, since

it considers the two most *different* poets in order to assess how *similar* the two groups are. The final dendrogram, however, usually does look like poets get split into several groups, and cuts can be meaningfully made. Finally, average linkage lacks the advantage that two actual poets are being compared; on the other hand, in our particular case, it does make sense to consider that the average of the group represents the group. Furthermore, the groups generated by average linkage are as interesting as those generated by complete linkage, but are easier to interpret. In the rest of this article, we will only consider average linkage.

4 Experiments

4.1 One language, many poets

TAB. 7 shows the rhythmic signature for 10-syllable verses in Portuguese, according to the poet who wrote them. Only poets with at least 2,500 such verses in our corpus were included; there were 15 poets that met this condition. By establishing this minimum amount, TAB. 5 tells us we can expect figures to be off by about 1.6%, so that care must be taken when interpreting smaller proportions.

A few figures immediately stand out from TAB. 7. Tomás Antônio Gonzaga employs pattern 2-6-8-10 with a much higher frequency, 24.2%, than any other poet; in fact, he uses them at more than twice the frequency of the second poet, José Teixeira, who uses it with a frequency of 10.6%. Looking at patterns with a stressed 4th syllable and unstressed 6th syllable (the so-called pure Sapphic type), we can see how this type is not much favored by Camões and it is consistently avoided by José Teixeira. Pattern 2-4-6-10 is the most common among 10 poets out of the total 15; pattern 3-6-10 was preferred by 3 poets; José Teixeira is alone in favoring 2-6-10; and Gonzaga, as already stated, has a strong penchant for 2-6-8-10 verses.

The Gini index of the 15 poets is given by TAB. 8. Poet Augusto dos Anjos has the lowest coefficient, at 0.374, and Gonzaga the highest, at 0.642. This means that the frequency of usage of the different rhythmic patterns is more evenly distributed in the case of Augusto dos Anjos, whereas Gonzaga uses a few meters with a high frequency; which can indeed be ascertained by manually examining TAB. 7. It is also worth noting that proximity in the table does not at all imply temporal proximity; in fact, both the oldest poet, Camões, and the youngest, José Teixeira, follow each other in the table, with coefficients of 0.524 and 0.580. Sá de Meneses, who follows Camões temporally and whose style was largely based on that of Camões, is far from his master in the table, with an index of 0.482.

FIG. 1 shows a dendrogram created from the rhythmic signatures. By following the labels placed on the dendrogram, one can derive interpretations for the groupings generated. At **A**, there is a split between Gonzaga, with his peculiar usage of the 2-6-8-10 rhythm. At **B**, the group of three poets have higher usages of patterns 3-6-10 and 2-6-10; in fact, the lowest usage of such patterns among the three poets is still higher than among the poets in the other group. At **C**, one can tell Teixeira apart by

	L. de Camões	Sá de Meneses	G. de Matos	S. R. Durão	C. M. da Costa	T. A. Gonzaga	G. de Magalhães	X. Pinheiro	F. Varela	D. Silveira	A. Figueredo	B. Tigre	A. dos Anjos	C. A. Nunes	J. Teixeira
2-4-6-10	15.2	12.3	13.4	11.2	12.7	9.0	8.2	16.5	7.3	15.3	12.8	11.3	10.1	7.7	11.8
3-6-10	10.3	12.0	12.2	10.3	9.6	6.5	11.9	9.7	14.5	5.9	8.8	9.6	9.9	14.1	13.9
2-6-10	9.0	9.7	9.9	7.1	8.7	13.1	8.4	6.8	11.3	5.2	4.6	5.6	8.8	13.2	16.3
2-4-6-8-10	11.1	9.6	8.5	9.1	10.3	13.4	7.2	11.3	4.1	9.8	8.6	9.7	6.9	5.1	8.1
2-6-8-10	7.8	7.1	5.9	7.6	8.1	24.2	7.1	5.3	9.4	3.2	2.5	5.2	3.6	9.6	10.6
3-6-8-10	7.7	8.2	7.7	8.5	8.9	10.4	8.2	6.1	9.0	4.3	4.2	7.2	3.4	7.3	9.1
1-3-6-10	6.2	5.2	8.5	7.7	6.5	2.9	10.2	7.2	6.1	6.1	7.9	7.9	7.0	6.8	7.0
1-4-6-10	7.9	5.8	7.7	6.1	6.3	1.9	6.7	7.5	7.3	11.4	10.7	8.3	7.5	5.1	4.0
2-4-8-10	1.1	6.6	3.5	8.1	5.9	3.8	5.8	6.7	6.5	11.8	9.3	6.9	7.8	4.6	0.2
1-3-6-8-10	4.5	3.9	6.4	5.8	5.3	6.5	7.8	4.8	4.2	3.4	4.0	6.0	2.8	3.5	4.4
1-4-6-8-10	5.0	3.9	4.2	4.6	4.9	2.2	5.1	5.3	3.3	5.9	6.8	6.5	4.7	2.7	3.1
4-6-10	6.2	5.5	4.8	3.6	4.2	1.8	3.1	4.4	5.5	4.6	4.8	4.1	4.8	5.3	4.6
1-4-8-10	0.5	2.3	1.6	3.6	3.1	1.1	4.6	3.6	4.6	7.5	6.3	4.3	6.2	3.5	0.1
4-6-8-10	4.2	3.6	2.3	2.9	2.7	1.8	2.9	3.1	2.5	2.2	2.6	2.7	2.7	3.5	2.7
4-8-10	0.4	1.9	0.8	2.2	1.0	0.3	0.8	1.0	1.3	1.7	3.1	1.8	3.7	3.1	0.1
1-6-10	1.3	1.1	1.2	0.4	0.7	0.5	0.7	0.3	1.7	0.7	0.8	0.7	3.9	2.1	1.8
1-6-8-10	1.0	0.8	1.0	0.8	0.9	0.5	1.1	0.2	1.3	0.4	0.3	0.6	1.3	1.4	1.2
6-10	0.3	0.3	0.2	0.0	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1	2.3	0.5	0.5
2-4-10	0.0	0.1	0.1	0.1	0.0			0.0	0.0	0.2	0.8	0.7	0.9	0.2	0.0
6-8-10	0.3	0.2	0.1	0.1	0.1	0.2	0.1	0.0	0.1	0.0	0.1	0.1	0.6	0.2	0.5
1-4-10	0.0	0.0	0.1	0.0				0.1		0.2	0.7	0.4	0.6	0.2	0.0
4-10	0.0	0.0	0.0	0.0				0.0		0.0	0.2	0.2	0.7	0.2	0.0

TAB. 7: Rhythmic signatures for poets who wrote at least 2,500 10-syllable verses in Portuguese

A. dos Anjos	B. Tigre	S. R. Durão	C. A. Nunes	G. de Magalhães	A. Figueredo	Sá de Meneses	C. M. da Costa	F. Varela	G. de Matos	X. Pinheiro	D. Silveira	L. de Camões	J. Teixeira	T. A. Gonzaga
0.374	0.442	0.461	0.462	0.464	0.467	0.482	0.483	0.487	0.506	0.506	0.515	0.524	0.580	0.642

TAB. 8: Gini coefficients for rhythmic signatures in TAB. 7, from lowest to highest

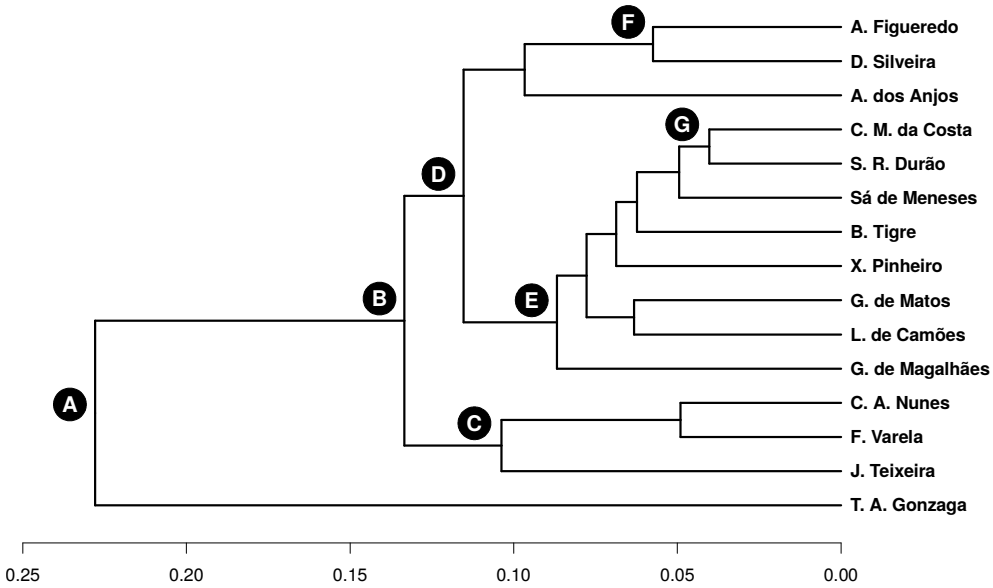

FIG. 1: Dendrogram for rhythmic signatures of TAB. 7

his extremely low usage of patterns such as 2-4-8-10 and 1-4-8-10 (pure Sapphic patterns). At **D**, we can separate the three poets by looking at the patterns 2-6-8-10 and 3-6-8-10: the three poets use them with a lower frequency than the other group, such that their maximum frequency is still lower than the other group's minimum frequency. At **E**, one can easily tell Magalhães apart by his low usage of the pattern 2-4-6-10. Similar reasonings can usually be found for all such decision points in dendrograms.

One cannot interpret the clusters in the dendrogram in terms of the times the poets lived in. While one can find groups such as **F**, which contains two poets that were born ten years apart and lived in the same town, and **G**, which contains two poets born seven years apart, these are exceptions. A poet's rhythmic signature does not depend on a place and time, but rather on style.

FIG. 2 shows a dendrogram created from 10-syllable verses collected from individual cantos of all eight epic poems written in Portuguese. Only cantos with at least 250 verses were included. This figure shows that the rhythmic signature of a poet remains more or less constant in a given epic poem. There are whole works that are completely contained in their own branch; such is the case of *Os Lusíadas*, *Bromilíadas*, *Anchieta*, *Os Timbiras, Malaca Conquistada*. *Famagusta* is contained in its own branch, except that the second canto of *Os Brasileidas* ends up in the mix. The remaining three epics, *Vila Rica*, *Caramuru* and *O Uraguai* are not well separated, although certain patterns can be spotted; indeed, the authors of the first two epics can be seen close together in FIG. 1, marked with label G. The third poet is not found in that figure because his epic poem fell short of the 2,500-verse threshold.

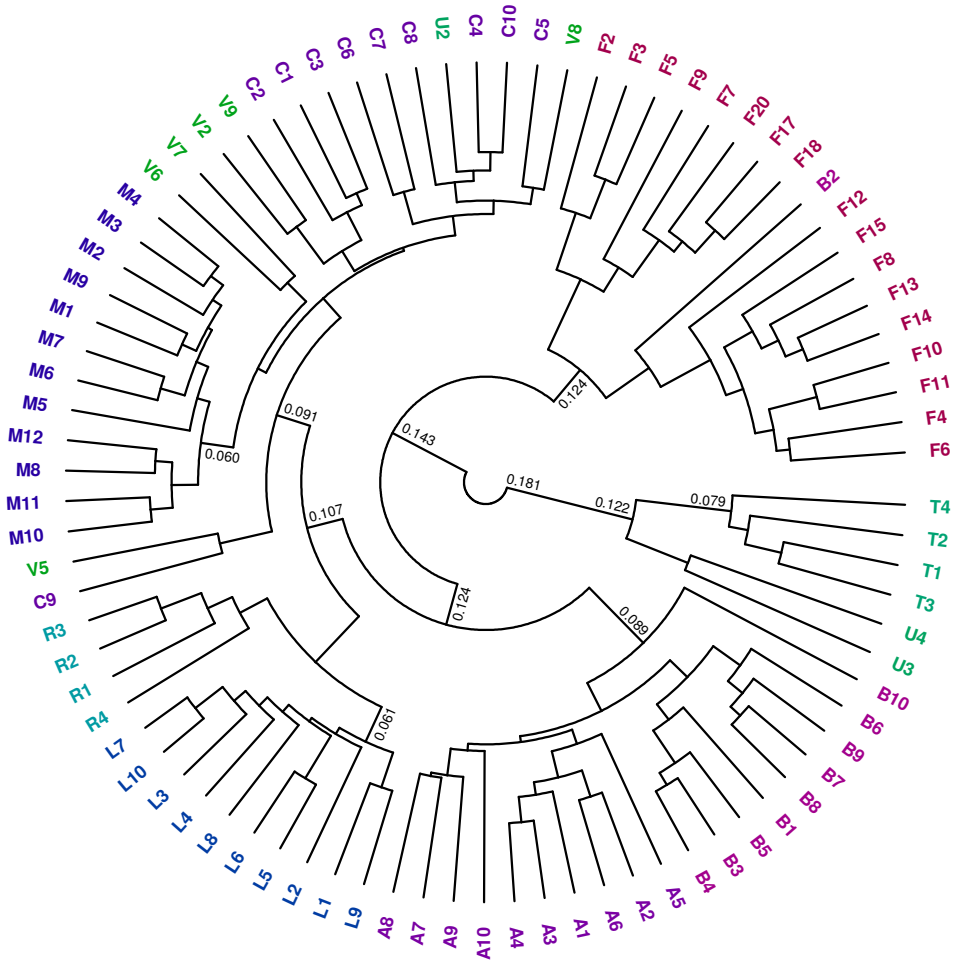

FIG. 2: Dendrogram for the rhythmic signatures of the individual cantos of 10 epic poems in Portuguese

TAB. 9 shows the rhythmic signature of poets with at least 1,000 7-syllable verses. TAB. 10 shows their Gini index and FIG. 3 shows the corresponding dendrogram. By comparing these results with those obtained with 10-syllable verses, we can immediately see that proximity in the dendrogram for one meter does not necessarily imply proximity in another meter: whereas FIG. 1 shows Delminda Silveira and Araújo Figueredo close together, that is no longer the case in FIG. 3. By examining TAB. 9, it becomes clear that Delminda Silveira has her own style when it comes to 7-syllable verses: pattern 2-4-7 is preferred among all poets, but she uses it with a far higher frequency (31.4%) than other poets, who use it at most with 21.4% frequency. By looking at the Gini coefficients of TAB. 10, it seems that the inequality of usage of patterns does not transfer among meters; there does not seem to be any relation between the Gini indices of TAB. 8 and TAB. 10.

	2-4-7	2-5-7	1-4-7	3-7	3-5-7	1-3-7	4-7	1-3-5-7	2-7	1-5-7	5-7	1-7	7
G. de Matos	18.8	12.8	12.0	11.8	9.8	8.1	8.3	6.6	6.0	3.3	1.7	0.7	0.1
D. Silveira	31.4	12.2	13.5	7.4	5.3	10.0	7.0	6.1	4.9	1.6	0.7	0.1	
A. Figueredo	17.1	13.3	9.5	12.6	10.1	11.4	9.8	9.1	3.9	2.1	1.0	0.1	
Bilac, Passos	21.4	15.0	13.4	7.7	8.3	9.3	5.5	9.1	5.8	2.8	1.1	0.7	0.1
B. Tigre	21.0	16.3	13.4	7.8	9.1	8.7	7.2	8.5	4.3	2.5	1.2	0.1	0.0

TAB. 9: Rhythmic signatures for poets who wrote at least 1,000 7-syllable verses in Portuguese

Poet	Gini index
G. de Matos	0.383
A. Figueredo	0.395
Bilac, Passos	0.428
B. Tigre	0.439
D. Silveira	0.513

TAB. 10: Gini coefficients for rhythmic signatures in TAB. 9, from lowest to highest

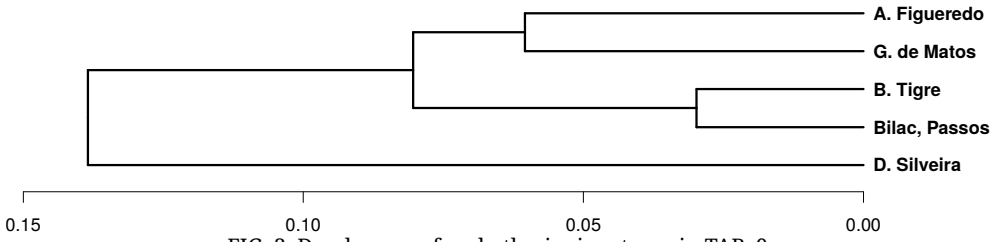

FIG. 3: Dendrogram for rhythmic signatures in TAB. 9

4.2 One poet, many books

Here we examine the 10-syllable verses contained in many works published by Brazilian poet Bastos Tigre, between 1902 and 1955. This writer has produced mainly humorous verses, for only 5 in a set of 22 books could be considered as non-humorous works (nevertheless, some poems in these five books are still clearly humorous). In fact, the analysis we propose here would allow us to compare this genre of poetry to the so-called serious poems which were (and always are) the mainstream of Brazilian poetry. Only works with at least 500 such verses were considered, so not allow too much uncertainty to leak into the results. FIG. 4 shows the dendrogram produced from the rhythmic signatures, and TAB. 11 the Gini indices.

It should first be noted that the poem *Bromilíadas* is a parody of Camões' *Os Lusíadas*. This was already clear in FIG. 2, and here also it is apparent that this work by Bastos Tigre is not like the remainder of his works: he was successful in imitating Camões in his style, to the point that even his rhythmic signature became much like that of the

older poet. Thus, in FIG. 4 *Bromilíadas* is set apart from the rest of his work, and the Gini index of this epic is also clearly distinct from the others in TAB. 11.

FIG. 4 does seem to imply certain temporal connections, such as the group that contains books published between 1913 and 1922. In general, however, Bastos' style seems to be disconnected from time: there are probably other variables (such as theme and tone) or even randomness influencing the results. His later works often include large portions of already published material—which were excluded from our analyses; it remains to be understood how his tendency towards reuse has affected, if it all, his rhythmic signatures.

Year	Work	Gini index
1905	Versos Perversos	0.408
1922	Fonte da Carioca	0.431
1902	Saguão da Posteridade	0.434
1919	Bolhas de Sabão	0.446
1955	Sol de Inverno	0.447
1913	Moinhos de Vento	0.452
1935	Entardecer	0.459
1922	Bromilíadas	0.525

TAB. 11: Gini coefficients for books published by Bastos Tigre, from lowest to highest; only books or poems with at least 500 10-syllable verses are included

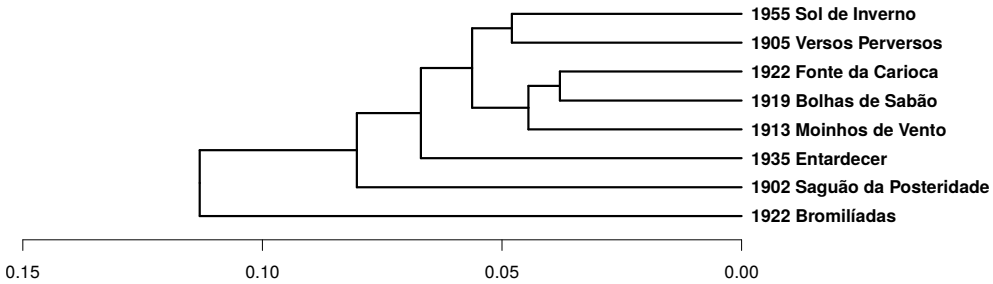

FIG. 4: Dendrogram for the rhythmic signatures of Bastos Tigre's works with at least 500 10-syllable verses

4.3 Two languages, many poets

TAB. 12 shows the rhythmic signatures calculated from all 10-syllable verses in Portuguese and Spanish. In general, they look similar, although clear differences do exist, such as the Spanish preference for pattern 2-6-10 (13.5% versus 9.2%) and the preference of Portuguese for pattern 2-4-6-8-10 (9.1% versus 6.7%). The Gini indices are very similar: 0.460 for Portuguese, 0.465 for Spanish.

	2-4-6-10	3-6-10	2-6-10	2-4-6-8-10	2-6-8-10	1-4-6-10	2-4-8-10	3-6-8-10	1-3-6-10	4-6-10	1-4-6-8-10	1-3-6-8-10	1-4-8-10	4-6-8-10	4-8-10	1-6-10	1-6-8-10	6-10	6-8-10	2-4-10	1-4-10	4-10
Pt.	11.9	10.8	9.2	9.1	7.7	6.7	5.9	7.5	6.7	4.5	4.6	4.8	3.3	2.9	1.5	1.1	0.8	0.3	0.2	0.2	0.1	0.1
Sp.	11.3	11.9	13.5	6.7	6.6	6.5	7.1	5.2	5.4	5.9	3.5	2.7	4.2	2.7	2.5	2.0	1.0	0.5	0.3	0.2	0.1	0.1

TAB. 12: Rhythmic signatures for Portuguese and Spanish; all 10-syllable verses are included

Language	Gini index
Portuguese	0.460
Spanish	0.465

TAB. 13: Gini coefficients for the rhythmic signatures of TAB. 12

Things are also interesting when we look at individual poets. We now consider all poets who wrote at least 1,000 10-syllable verses: 19 who wrote in Spanish, 17 who wrote in Portuguese. Tables for the rhythmic signatures and the Gini coefficients are not shown due to size constraints, but a few numbers are worth mentioning.

Gonzaga has found a match to his single-mindedness: Centenera, the author of the epic *La Argentina*, uses pattern 2-6-10 in a whopping 27.1% of his verses and his Gini coefficient is 0.706; compare these figures to Gonzaga's 24.2% usage of pattern 2-6-8-10 and Gini coefficient of 0.642. Like Teixeira, it seems Centenera's intention was to avoid Sapphic verses altogether: patterns that include the 4th (but not the 6th) syllable are very infrequent in both poets, with a maximum of 0.2% in Centenera for pattern 2-4-8-10, against the global average of 7.6%. Augusto dos Anjos still holds the lowest Gini index: 0.374.

The dendrogram for the 36 poets can be seen in FIG. 5. It is clear that rhythmic signatures cannot distinguish between Portuguese and Spanish, but some patterns are nonetheless visible. First, all poets not included in group **A** have something peculiar about their rhythmic signature, regardless of language. Centenera has by far the highest frequency of 2-6-10 verses: 27.1%. Gonzaga, by far that of 2-6-8-10 verses: 24.2%. Francisco de Borja uses the pattern 2-4-8-10 with a frequency of 24.6%, far above the second highest usage, 16.0%. Luis de Ulloa Pereira has simultaneously the highest usage of 3-6-10 verses, 16.9%, and the lowest of 2-4-6-10, 6.2%. Gonçalves Dias stands apart for the highest frequency of pattern 2-4-6-8-10, at 14.5%, and the lowest of pattern 3-6-10, at 4.4%. Finally, José Teixeira has the lowest frequency of 2-4-8-10 verses, along with Centenera; but, unlike the latter, his Gini index is not as high: while Centenera tops the list with a coefficient of 0.706, Teixeira's coefficient is 0.580.

In general, group **B** contains mostly (80%) poets who wrote in Portuguese. Group **C**, on the other hand, has predominantly (79%) poets who wrote in Spanish. Group **D** is mixed. As with previous experiments, time does not seem to be of paramount importance for rhythmic signatures.

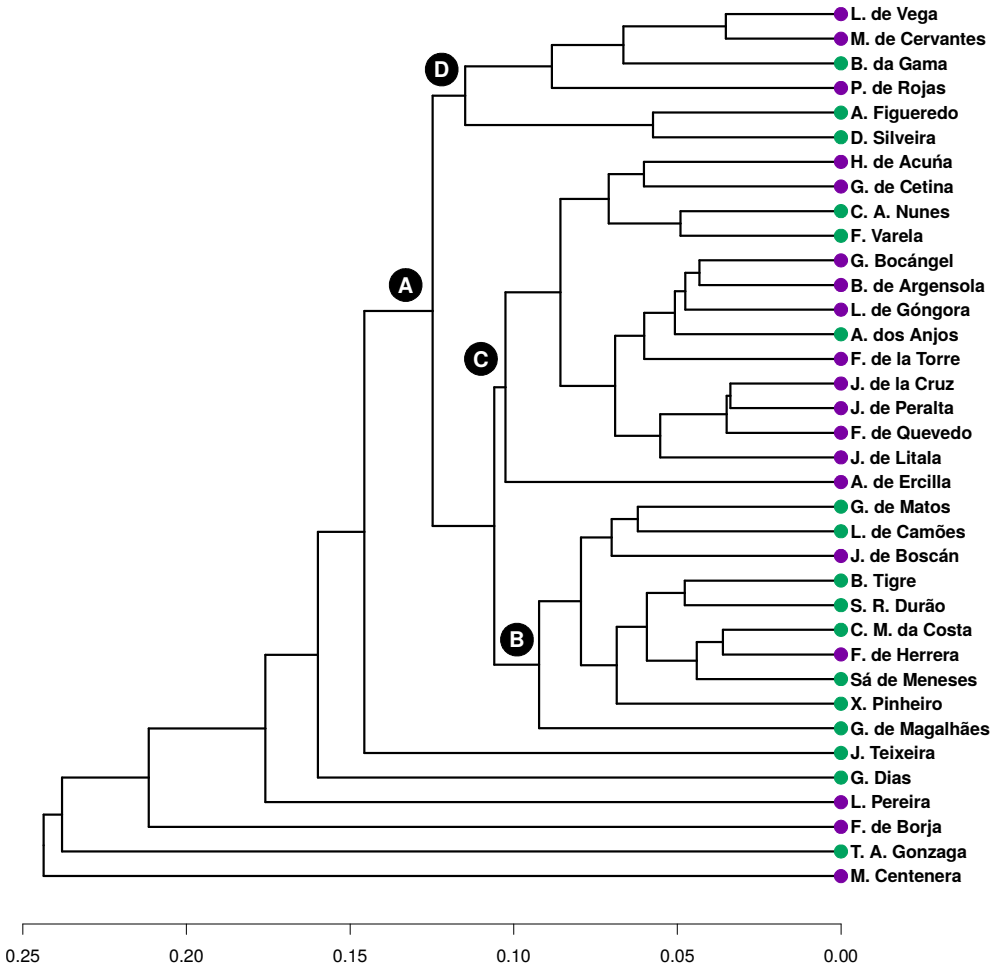

FIG. 5: Dendrogram for the rhythmic signatures of poets who wrote at least 1,000 10-syllable verses. ● = Portuguese, ● = Spanish

5 Discussion

It is, above all else, clear that, even though the rhythmic variability of a certain poet does depend partially on the times he lived in, we cannot distinguish rhythmic trends that would take us from one generation of poets to the next. There is no such progression in the usage of metric and rhythmic elements in poetry. In our case, the fact that 20th century poet Augusto dos Anjos presents more variation in rhythmic patterns that 18th century Tomás Antônio Gonzaga is nothing but a coincidence. Camões (16th century) and José Teixeira (21st century) present similar Gini coefficients, although they have published their works more than four centuries apart. Rhythmic richness depends primarily on a set of elements, such as the literary genre, the subject matter

of the poem, and, perhaps more importantly, the poet's own style. This last one—the style—establishes a sort of dialectical relationship with the habits of a literary period: an author's particular style and the period's general literary style set up some sort of mutual influence, such as can be seen in the almost exclusive usage of the 4-7-10 rhythmic pattern in Provence (verses with this pattern, not coincidentally, are called *decassílabos provençais* in Portuguese). That is why we can propose a rhythmic signature, an element which poets use, likely unconsciously, when making their verses.

Some interesting questions arise when we focus on the differences among the rhythmic signatures of the poets we analyzed. In the dendrogram shown in FIG. 1, there are manifest and unsurprising proximities, such as that between Araújo Figueredo and Delminda Silveira: they are both from the same Brazilian region (State of Santa Catarina) and belong to the same literary generation, two conditions that would explain their proximity. On the other hand, Bastos Tigre (early 20th century), is close to romantic poet and translator Xavier Pinheiro (early 19th) and to a group that includes baroque poet Sá de Meneses (17th century). If we had also analyzed Brazilian Parnassians, then probably Bastos Tigre would be much closer to them.

When we examine the cantos of epic poems, we can find an evident homogeneity in the case of Camões' *Os Lusíadas*, Bastos Tigre's *Bromilíadas*, Varela's *Anchieta,* Gonçalves Dias' *Os Timbiras* and José Teixeira's *Famagusta.* This could mean that it did not take very long for these poets to write their epics, which made their style more uniform throughout all cantos. Some differences might appear when cantos were written over a longer period of time, or if they were modified a long time after the original composition. Thus, analyzing style variations over a period of time could bring about important information concerning the change of an author's style, which would be useful in determining when a work whose time of composition is unknown was written.

It is important, however, to highlight the case of Bastos Tigre's *Bromilíadas*: it is a direct parody, which means that he followed very closely Camões' epic poem. It makes sense that, if there is no large rhythmic variation in the original, the same thing will happen to the parody. Furthermore, as one may observe in FIG. 2, FIG. 4 and TAB. 11, the style of the *Bromilíadas* is much closer to the work of Camões than any other book by Bastos Tigre, which demonstrates the poet's very accurate effort in following *Os Lusíadas*.

It is also relevant to investigate the rhythmic signatures of TAB. 9, which were calculated from the 7-syllable verses of five Brazilian poets. The fourth "poet" is, actually, a hybrid, for it was a duo (Olavo Bilac and Guimarães Passos) who wrote the book, *Pimentões,* where the 7-syllable verses were taken from; in this case, evidently, one does not obtain the style of an actual poet, but something like a mixed style, though it is certainly close to Bilac's style—especially true when we consider the enormous influence he had on his literary generation, including Guimarães Passos himself. It is also important to stress that 7-syllable verses permit a good amount of variation in the set of actual (not theoretical) rhythmic patterns: 13 possibilities, which is not as small as the number of patterns in 4-, 5- or 6-syllable verses. The result is expressed by the simple dendrogram of FIG. 3. One would expect that both Bastos Tigre and

Bilac–Passos were closer to Gregório de Matos, since all of them wrote satirical poems; however, the poet who is closer to Matos is 19th century poet Araújo Figueredo, who did not write satirical verses. Probably the usage of popular versification was, in this case, a more important factor in determining the resemblance of style than the subject matter or the strategy of the poems.

TAB. 12 shows the rhythmic signatures calculated from a large number of verses for two languages, Portuguese and Spanish. Comparing the numbers, we can readily see that some rhythmic patterns are used in very close proportions: 2-4-6-10, 1-4-6-10, 4-6-8-10, 1-6-8-10, 6-8-10, 2-4-10, 1-4-10, 6-10, 4-10. Hence, among 22 rhythmic patterns, 9 (41%) are very similarly used by poets of both groups. This means that there is some difference in the usage of almost 60% of the rhythmic patterns. However, not all rhythmic patterns have the same importance, for they present very different frequencies; the more frequently used patterns are more important when examining a rhythmic signature. If we establish that the difference between two frequencies are relevant if they both are at least 5% and the ratio of the smaller to the larger is at most 0.75, then we are left with three pattern whose frequencies are relevantly different:

- 2-6-10: Portuguese 9.2%, Spanish 13.5%, ratio of 0.68;
- 2-4-6-8-10: Portuguese 9.1%, Spanish 6.7%, ratio of 0.74;
- 3-6-8-10: Portuguese 7.5%, Spanish 5.2%, ratio of 0.69.

Taken individually, poets present clear differences concerning their rhythmic signatures. As we said earlier, the fact that a poet writes in Portuguese or Spanish does not entail a major rhythmic distinction between them, just as their times do not either. Brazilian poet Augusto dos Anjos, for instance, is closer to Spanish poets from the 16th and 17th centuries than to his contemporary Bastos Tigre, according to FIG. 5. Not coincidentally, many literary scholars say that Augusto dos Anjos' poetry is similar to that of the... baroque! It is also interesting that, in the same figure, baroque poets Góngora and Lope de Vega are distant from one another. To put it succinctly, individual style is much more important than other elements, and this fact allows us to compare poets from all ages. We can claim that their styles, translated into their preferences for rhythmic patterns, build up a veritable rhythmic signature.

6 Conclusion

Rhythmic signature aids in characterizing the style of a poet or, more generally, of a set of verses. The role of automatic scansion tools, in this context, is paramount: in this article we examined more than 250,000 verses, which would take a prohibitively long time to scan and annotate manually. Aoidos, our scansion tool, is capable of providing rhythmic signatures for arbitrary sets of verses, and was used for producing the results shown in this article.

We believe that automatic analysis of poetry in general and automatically-calculated rhythmic signatures in particular are a useful tool for providing objective information to experts. We hope that this sort of information can foster discussion and pro-

mote a more in-depth understanding of poetry, both past and present. The kind of discussion contained in Sect. 5 is a sample of what can now be achieved when the rhythmic patterns of thousands of verses and dozens of poets are analyzed, tabulated and graphically viewed.

There is still plenty of work ahead. As one example, the analyses produced by Navarro-Colorado (2016) on his Siglo de Oro corpus are different from the one presented here. The results are more or less compatible, but are different in the details. In particular, the way rhythmic patterns are computed by Navarro-Colorado is not the same as ours: whereas he talks about patterns such as 2-3-6-10, our tool never produces such patterns. Indeed, research is needed in order to quantify and understand how both humans and machines see rhythmic patterns.

Acknowledgment

We thank Samanta Maia for producing and sharing the digital editions of the works of Bastos Tigre. Dendrograms in this article were produced with R; packages `dendextend` (Galili 2015) and `circlize` (Gu et al. 2014) have been used.

References

Beaudouin, V. – Yvon, F. (1996). The metrometer: A tool for analysing French verse. *Literary and Linguistic Computing*, *11* (1), 23–31.

Chociay, R. (1994). A identidade formal do decassílabo em “O Uraguai”. *Revista de Letras*, *34*, 229–243.

Delente, É. – Renault, R. (2015). Projet anamètre: Le calcul du mètre des vers complexes. *Langages*, *3* (199), 125–148.

Galili, T. (2015). dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, *31*, 3718–3720.

Gu, Z. – Gu, L. – Eils, R. – Schlesner, M. – Brors, B. (2014). *circlize* implements and enhances circular visualization in R. *Bioinformatics*, *30*, 2811–2812.

Mittmann, A. – Luiz dos Santos, A. (2018). The rhythm of epic verse in Portuguese from the 16th to the 21st century. In *14èmes Journées Internationales d’Analyse Statistique des Données Textuelles* (514–521).

Mittmann, A. – Maia, S. R. (2017). Análise comparativa entre escansões manual e automática dos versos de Gregório de Matos. *Texto Digital*, *13* (1), 157–179.

Mittmann, A. – von Wangenheim, A. – Luiz dos Santos, A. (2016). Aoidos: A system for the automatic scansion of poetry written in Portuguese, 2016. In *17th International Conference on Intelligent Text Processing and Computational Linguistics*, *2*, 611–628.

Moretti, F. (2013). *Distant reading*. London: Verso.

Navarro-Colorado, B. (2016). Hacia un análisis distante del endecasílabo áureo: Patrones métricos, frecuencias y evolución histórica. *Rhythmica*, (14), 89–118.

Navarro-Colorado, B. – Lafoz, M. R. – Sánchez, N. (2016). Metrical annotation of a large corpus of Spanish sonnets: Representation, scansion and evaluation. In *9th International Conference on Language Resources and Evaluation*, 4360–4364.