Computational identification of phonological parallelisms in Brazilian literary prose

Leonardo Lima[1][0000-0002-2414-5672]⋆, João Queiroz[2][0000-0001-6978-4446], and Angelo Loula[1][0000-0001-7802-1731]

[1] Universidade Estadual de Feira de Santana, Feira de Santana BA, Brazil
leo.os.lima@gmail.com angelocl@uefs.br
[2] Universidade Federal de Juiz de Fora, Juiz de Fora RJ, Brazil
queirozj@gmail.com

Abstract. The identification of parallel patterns of phonological structures in literary prose is still an unexplored phenomenon in Literary Studies and Digital Humanities. We propose a computational method for searching and identification of phonological parallelisms in Brazilian Portuguese literary prose that outputs sequences of repeated syllables, based on two parameters related to repetition closeness and minimum number of repeated occurrences. Preliminary results from processing three literary works are presented, with quantitative measures and a co-location perspective of sequences.

Keywords: Computing · Literary Prose · Phonological Parallelisms .

1 Introduction

It is well known that the “structure of poetry is that of a continuous parallelism”[7], at several organization levels: grammatical, semantic, syntactic, phonological, rhythmic, prosodic, even typographic. This phenomenon has no correspondence in prose. But, as Jakobson[7] argues, “literary prose lies between poetry as such and the common, practical language of communication, and it must not be forgotten that it is incomparably more difficult to analyze an intermediate phenomenon, transition, than studying extreme phenomena.” In literary prose, several forms of parallelism are distributed in many levels, allowing the identification of patterns of phonological structures under the seemingly uniform surface of writing, a surprising and still little explored phenomenon in Digital Humanities for Literary Studies.

We are especially interested in the occurrence of phonological parallelisms in literary prose of Portuguese language. Usually defined as the repetition of consonant sound close enough to affect the listening of the reader[2], alliterations are among the best known examples of phonological parallelisms (others are assonance, consonance, rhyme). Phonological parallelism in prose has been studied in previous works in Literary Studies, outside Natural Language Processing and

Digital Humanities, but it is an open challenge to computationally identify such phenomena and also to identify the behavioral patterns of phonological parallelisms at various observational scales, from one literary work to many works, by the same author, group of authors, or aesthetic movements.

Here we propose a computational method for searching and identification of repeated (phoneme) syllables in Brazilian Portuguese literary prose. The computational algorithms are guided by parameters that influence constraints in the retrieval of sequences of repeated syllables in the provided text. We present quantitative measures to describe preliminary results obtained from three works − *Os Sertões* (*Rebellion in the Backlands*, 1902), by Euclides da Cunha, *Triste fim de Policarpo Quaresma* (1915), by Lima Barreto, and *Macunaima* (1928), by Mario de Andrade − and also exhibit a co-location perspective of results.

There have been previous works in Digital Humanities for computational identification of phonological patterns in literary works with close and distant reading approaches[1, 3, 8, 9], but focused on poetry. Such phonological patterns have also been used for authorship attribution of literary work[5, 6]. As far as we know, the computational search for phonological parallelism in prose was not studied before.

2 Methodology

The computational identification of phonological parallelisms concerns the search for repetitions of sounds along a literary work. Among possible sound elements, for preliminary experiments, we chose the syllable as the unit of interest and define this computational task as the search for sequences of repeated (phoneme) syllables inside a sliding window containing a given number of words (window size), filtering sequences above a given length. Window size sets a minimum closeness context as only syllables that co-occur inside the text window are considered a repetition. Sequence length defines a minimum number of repetitions for a sequence of the same syllable to be considered noteworthy. These parameters guide a phoneme pairs index construction algorithm and a sequence construction and filtering algorithm.

The first step for the computational process is preprocessing the prose text to tokenize words followed by a syllabification and grapheme to phoneme conversion for Brazilian Portuguese, using the UFPAT2S tool[4]. It is based on the set of rules for the conversion and determination of stressed syllables [10], achieving an accuracy of 97.44% and 98.58%, respectively, for the grapheme-phone convert and the stress determination algorithm[4].

After conversion, the phonetic version of the original text is obtained, with words written as phoneme syllables (referred just as syllable from now on) in SAMPA alphabet with syllabic divisions, for example the word ⟨madeira⟩ (wood in Portuguese) is converted to /ma-'dej-ra/. The phonetic version of the text is then processed by algorithms to obtain the final syllable sequences of interest.

The computational task is algorithmically divided in two steps. The first algorithm runs through the text with a sliding window of size S (corresponding to S words) to find pairs of repeated syllables, constructing an index of same syllable occurrences. This algorithm considers every syllable in the first word in the window and looks for the first next occurrence of the same syllable inside the window. After that, the window slides, dropping the first word and adding the next word in text at the end of the window, and repeating the same search for syllables for the new first word. The window continues sliding until it reaches the end of the text. The output of this algorithm is an index of pairs of occurrences of same syllables along the text, where each entry is an occurrence (syllable-position tuple) pointing to the next occurrence of the same syllable.

The following example illustrates the process of identifying pairs of repeated syllables, for a window size of 3 words. The input sentence is converted to phonological form with separated syllables. The output indicates the repeated syllables (underlined) and the corresponding syllable-position tuples.

Input: sob o mando do coronel Tamarindo, separada
da esquerda, dirigida pelo capitão Felipe Simões
⇓
Output: 'sob u 'ma∼-du 'du ko-ro-'nEw ta-ma-'ri∼-du se-pa-'ra-da
'da es-'keX-da dZi-ri-'Zi-da 'pe-lu ka-pi-'ta w fe-'li-pi
(du,4),(du,5),(du,12),(da,16),(da,17),(da,20),(da,24)

The second algorithm uses the index of pairs of syllable occurrences to construct full sequences and then filters only those that are above the minimum sequence parameter length L. Head occurrences, which start a sequence, are identified as those that are not next occurrences to any other occurrence. From the head occurrence, the algorithm constructs a sequence similarly to a linked list, traversing the occurrences pointed as next ones until a next occurrence is not found, thus ending the sequence. Only sequences with length equal or greater than L are outputted at the end of this process. The final result of the computational search is a list of sequences of repeated syllables along with the position of each of them, based on the parameters provided.

Using the output of the previous example as input, the following is an example of the output obtained after constructing full sequences with minimum length of 4 occurrences. There were two possible sequences, with syllable /du/ and syllable /da/, but the output highlights that only the second one was considered, as the first has length below minimum size.

Input: 'sob u 'ma∼-du 'du ko-ro-'nEw ta-ma-'ri∼-du se-pa-'ra-da
'da es-'keX-da dZi-ri-'Zi-da 'pe-lu ka-pi-'ta w fe-'li-pi
(du,4),(du,5),(du,12),(da,16),(da,17),(da,20),(da,24)
⇓
Output: sob o mando do coronel Tamarindo, separada
da esquerda, dirigida pelo capitão Felipe Simões
[(da,16),(da,17),(da,20),(da,24)]

3 Results

Initial results in our research were obtained by applying our computational search process to three classic works of Brazilian literary prose: *Macunaíma*, *Os Sertões*, and *Triste Fim de Policarpo Quaresma*. The selected works are some of the most representative literary experiments in the transition between pre-modernism and modernism in Brazil. The computational search for sequences of repeated syllables is guided by the parameters S and L. Both influence the number of sequences obtained as the final result, as shown in Table 1.

Table 1: Number of sequences of repeated syllables in literary prose books, for different values of window size S and minimum sequence length L. Absolute (abs) values and relative (rel) values, in relation to total number of words.

S	L	Macunaíma		Os Sertões		Triste Fim	
		abs	rel	abs	rel	abs	rel
2	3	200	0.46%	1829	1,15%	851	1.27%
2	4	39	0.09%	383	0,24%	188	0.28%
2	5	16	0.04%	100	0,06%	49	0.07%
3	3	449	1.03%	3354	2,11%	1525	2.28%
3	4	92	0.21%	923	0,58%	388	0.58%
3	5	38	0.09%	288	0,18%	124	0.19%
4	3	782	1.79%	4461	2,90%	2205	3.30%
4	4	179	0.41%	1572	0,99%	646	0.97%
4	5	64	0.15%	564	0,35%	236	0.35%
5	3	1170	2.68%	6671	4,19%	2836	4.24%
5	4	297	0.68%	2295	1,44%	912	1.36%
5	5	118	0.27%	925	0,58%	381	0.57%

Table 2: Number of sequences per repeated syllable in literary prose books, for S=4 and L=5 (gray line in Table 1). Only the 10 most frequent syllables (syl) are shown.

Macunaíma		Os Sertões		Triste Fim	
syl	abs	syl	abs	syl	abs
a	35	a	462	a	182
e	13	du	30	dZi	18
u	10	dZi	20	du	6
ma	7	u	10	tSi	4
ra	6	e	6	e	3
du	6	tSi	6	u	3
dZi	4	si	5	si	2
ku	4	ra	4	ra	2
da	4	da	4	ew	1
ta	2	ka	2	fo	1

Depending on parameter values, the computational method was able to identify different numbers of sequences of repeated syllables. As window size grows, more sequences are found, since occurrences would be farther away are considered. On the other hand, as minimum sequence length increases, less sequences are selected, once shorter ones are dropped. Comparing the literary works, there are distinct absolute numbers of sequences identified for the same parameters, mainly due the difference in total words in each one, with *Macunaíma* having 43.665 words, *Os Sertões*, 159.070 words and *Triste fim de Policarpo Quaresma*, 66.869 words. Nevertheless, the relative percentage of sequences reveals that *Os Sertões* and *Triste Fim de Policarpo Quaresma* have higher densities of phonological parallelisms.

Besides the total number of sequences in literary works, a quantitative analysis of distribution of distinct repeated syllables among sequences is shown in Table 2, considering the gray line from Table 1. This configuration involves parameters quite strict for sequences as only sequences with 5 syllable repetitions within 4 words are reported. There is a high concentration of sequences with certain syllables. The greatest number of sequences involve the phoneme /a/ as an isolated vowel syllable. Vowels are centers of syllables and often occur isolated in syllables, but /a/ has a much higher frequency than other vowels. /a/ can also be used as a singular definite article ⟨a⟩, but the other singular definite article, ⟨o⟩ /u/, does not have such high frequency. But there are also sequences with consonants among the most frequent ones and different sequences can overlap in the processed books.

Collocations of sequences occur in the book. The following excerpt is an example of such overlap of repeated syllable sequences, involving syllable /ma/ in word 11.595, syllable /u∼/ in word 11.597, and /ku/ in word 11.598, using different colors to highlight the three sequences.

> “Maanape não queria jogar o mano mesmo, pegou desesperado em seis caças duma vez um macuco um macaco um jacu uma jacutinga uma picota e uma pia-coça e atirou no chão gritando:”

The /ma/ and /u∼/ have length of 6 occurrences and /ku/ sequence has length of 5 occurrences, so both fulfil the length threshold of 5 repetitions. A higher L value would drop off some of these sequences. The window had size of 4 words, allowing an occurrence in a first word to be connected to a occurrence in the three words ahead. A smaller window size would break some of these connections, breaking or shortening sequences. Changing these two parameters allows analysis from different perspectives, more strictly or more loose.

4 Final Remarks

This is a work in progress in the computational search and identification of repeated syllable sequences in Brazilian Portuguese literary prose. Such sequences can be seen as parallelisms at phonological level, that can be analyzed according to diverse exploratory perspectives in Digital Humanities.

The proposed computational process is guided by two parameters that vary the final number of sequences obtained and further study is needed to characterize and compare the sequences obtained with different configurations. Other sound units (e.g. individual phonemes) and clusters, similarity criteria and tonicity constraints will be addressed in future work.

Preliminary results from computational identification of phonological parallelism evidence a dense presence of repeated sequences in literary prose. Nevertheless, more literary works will be evaluated, as well as ordinary prose as a control group, considering other authors and other literary periods. In addition, other descriptive perspectives of distant reading and close reading will be evaluated.

References

1. Abdul-Rahman, A., Lein, J., Coles, K., Maguire, E., Meyer, M., Wynne, M., Johnson, C.R., Trefethen, A., Chen, M.: Rule-based visual mappings–with a case study on poetry visualization. In: Computer Graphics Forum. vol. 32, pp. 381–390. Wiley Online Library (2013)
2. Adam, P.G., Cable, T.: Alliteration. In: Greene, R., Cushman, S., Cavanagh, C., Ramazani, J., Rouzer, P., Feinsod, H., Marno, D., Slessarev, A. (eds.) The Princeton Encyclopedia of Poetry and Poetics, p. 40–42. Princeton University Press (2012)
3. Benner, D.C.: The sounds of the psalter: Computational analysis of soundplay. Literary and Linguistic Computing **29**(3), 361–378 (2014)

4. Costa, E., Neto, N.: Free tools and resources for hmm-based brazilian portuguese speech synthesis. In: Simari, G.R., Fermé, E., Gutiérrez Segura, F., Rodríguez Melquiades, J.A. (eds.) Advances in Artificial Intelligence - IBERAMIA 2018. pp. 367–379. Springer International Publishing (2018)
5. Ivanov, L.: Using alliteration in authorship attribution of historical texts. In: International Conference on Text, Speech, and Dialogue. pp. 239–248. Springer (2016)
6. Ivanov, L.: Learning patterns of assonance for authorship attribution of historical texts. In: The Thirty-Second International Flairs Conference (2019)
7. Jakobson, R., Pomorska, K.: Dialogues. MIT Press (1988)
8. Kao, J., Jurafsky, D.: A computational analysis of style, affect, and imagery in contemporary poetry. In: Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature. pp. 8–17 (2012)
9. McCurdy, N., Srikumar, V., Meyer, M.: Rhymedesign: A tool for analyzing sonic devices in poetry. In: Proceedings of the Fourth Workshop on Computational Linguistics for Literature. pp. 12–22 (2015)
10. Silva, D.C., de Lima, A.A., Maia, R., Braga, D., de Moraes, J.F., de Moraes, J.A., Resende, F.G.: A rule-based grapheme-phone converter and stress determination for brazilian portuguese natural language processing. In: 2006 International Telecommunications Symposium. pp. 550–554. IEEE (2006)