UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE ENGENHARIA ELÉTRICA
ENGENHARIA ELÉTRICA

SAMANTHA TRANSFELD

DATA MINING E PROCESSAMENTO DIGITAL DE SINAIS PARA A ANÁLISE DE ESTRUTURAS POÉTICAS

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA
2018

SAMANTHA TRANSFELD

DATA MINING E PROCESSAMENTO DIGITAL DE SINAIS PARA ANÁLISE DE ESTRUTURAS POÉTICAS

Trabalho de Conclusão de Curso de Graduação, apresentado ao curso de Engenharia Elétrica do Departamento Acadêmico de Eletrotécnica (DAELT) da Universidade Tecnológica Federal do Paraná (UTFPR), como requisito parcial para obtenção do título de Engenheiro Eletricista.

Orientador: Professor Doutor Marcelo de Oliveira Rosa.

CURITIBA

2018

SAMANTHA TRANSFELD DA SILVA

Data Mining e Processamento Digital de Sinais para Análise de Estruturas Poéticas

Este Trabalho de Conclusão de Curso de Graduação foi julgado e aprovado como requisito parcial para a obtenção do Título de Engenheiro Eletricista, do curso de Engenharia Elétrica do Departamento Acadêmico de Eletrotécnica (DAELT) da Universidade Tecnológica Federal do Paraná (UTFPR).

Curitiba, 13 de junho de 2018.

_____

Prof. Antonio Carlos Pinho
Coordenador de Curso
Engenharia Elétrica

_____

Profa. Annemarlen Gehrke Castagna, Me.
Responsável pelos Trabalhos de Conclusão de Curso
de Engenharia Elétrica do DAELT

ORIENTAÇÃO

_____
Marcelo de Oliveira Rosa, Dr.
Universidade Tecnológica Federal do Paraná
Orientador

BANCA EXAMINADORA

_____
Glauber Gomes de Oliveira Brante, Dr.
Universidade Tecnológica Federal do Paraná

_____
Gustavo Nishida, Dr.
Universidade Tecnológica Federal do Paraná

_____
Marcelo de Oliveira Rosa, Dr.
Universidade Tecnológica Federal do Paraná

A folha de aprovação assinada encontra-se na Coordenação do Curso de Engenharia Elétrica.

Às pessoas maravilhosas que caminham comigo, pela inspiração e pelo carinho, meu sincero agradecimento.

RESUMO

DA SILVA, Samantha Transfeld. ***Data Mining* e Processamento Digital de Sinais para a Análise de Estruturas Poéticas.** 2018. 65f. Trabalho de Conclusão de Curso (Bacharelado em Engenharia Elétrica) – Universidade Tecnológica Federal do Paraná. Curitiba, 2018.

Este trabalho tem como objetivo validar a aplicação do processamento digital de sinais em estruturas literárias. Desta forma, a análise é validada baseando-se nos princípios da mineração de dados para a compreensão matemática de poemas na língua portuguesa. O reconhecimento de padrões é um campo em rápido desenvolvimento, que sustenta estudos em áreas tal qual a busca pela métrica proveniente da rima obtida em textos poéticos. A validação de tal característica é realizada com a aplicação da autocorrelação como ferramenta matemática. A partir desta análise, assume-se que os maiores valores de autocorrelação representem a maior similaridade do texto. Esta ferramenta possibilita a interpretação de sinais ao buscarmos por padrões de repetição em poemas. Assim, o melhor resultado da análise leva em consideração apenas o final de cada verso, onde assume-se que há a rima. Nestes casos, são obtidos os maiores valores de autocorrelação, ainda que tais valores não apresentem significativa relevância para a padronização do texto.

Palavras-chave: Autocorrelação. *Data Mining*. Reconhecimento de padrões. Poesia.

ABSTRACT

DA SILVA, Samantha Transfeld. **Data Mining and Digital Signal Processing for the Analysis of Poetic Structures.** 2018. 65f. Trabalho de Conclusão de Curso (Bacharelado em Engenharia Elétrica) – Universidade Tecnológica Federal do Paraná. Curitiba, 2018.

This work aims to validate the application of digital signal processing in literary structures. In this way, the analysis is validated based on the principles of data mining for the mathematical understanding of poems in the Portuguese language. Pattern recognition is a rapidly developing field that supports studies in areas such as the search for metrics from the rhyme obtained in poetic texts. The validation of such characteristic is accomplished with the application of autocorrelation as a mathematical tool. From this analysis, it is assumed that the higher values of autocorrelation represent the greater similarity of the text. This tool enables the interpretation of signals as we search for patterns of repetition in poems. Thus, the best result of the analysis considers only the end of each verse, where it is assumed that there is rhyme. In these cases, the highest values of autocorrelation are obtained, although these values do not present significant relevance for the standardization of the text.

Keywords: Autocorrelation. Data Mining. Pattern Recognition. Poetry.

Índice de Figuras

Índice de Tabelas

SUMÁRIO

1. INTRODUÇÃO

A comunicação é essencial para o desenvolvimento psíquico humano, introduzindo e modificando padrões culturais e comportamentais na sociedade. Assim, conceitos de cultura e comunicação são redefinidos constantemente, de acordo com valores intrinsecamente relacionados à evolução humana.

O presente capítulo abrange uma introdução à aplicação do processamento digital de sinais para a análise e identificação de estruturas literárias na língua brasileira. Explora-se, neste capítulo, a utilização de conhecimentos de engenharia para o avanço de tecnologias na compreensão de padrões comunicacionais.

1.1 TEMA

É notável o crescimento de sistemas de comunicação dentro da sociedade moderna, tornando o processo comunicacional cada vez mais veloz e eficiente. O desenvolvimento de novas tecnologias tem revolucionado a gama de recursos informativos, criando novas e mais robustas formas de comunicabilidade dispostas a favor do conhecimento.

Assim, como desenvolvimento de uma nova tecnologia, são estudadas informações referentes à padronização de estruturas literárias na língua brasileira com a utilização do processamento digital de sinais.

1.1.1 Delimitação do Tema

A remodelagem das formas de comunicação tradicionais deve-se, principalmente, à incorporação da internet na transmissão de dados. Sendo este considerado o maior exemplo de um sistema complexo projetado em grande escala, caracterizado por seu enorme grau de heterogeneidade (WILLINGER et al., 2006).

De tal modo, a diversificação dentro de tal sistema permite que haja um maior número de transformações na produção de informações, permitindo a interconectividade em diversos âmbitos, propiciando a expansão cultural e o dinamismo da comunicação cibernética.

Dentro do desenvolvimento digital, destaca-se a relevância da atribuição da tecnologia no âmbito social atual e sua utilização na formulação de discursos, propondo-se a ampliar o panorama da produção textual e a construção de linguagem no meio digital (RIBEIRO et al., 2012).

As possibilidades de recepção de conteúdo literário abrangem a temática poética, configuradas como um aditivo do sistema literário, alterando e prolongando o ciberespaço (MATIA, 2013). Isso posto, quando inserida no modelo eletrônico, revela-se muito da poesia tradicional, proporcionando elementos estéticos e literários a escritores e o estímulo abrangente à leitura àquele que lê.

A poesia, caracterizada pela marcação rítmica, utiliza padrões específicos para atingir efeitos estéticos. Tais padrões podem facilitar o reconhecimento e a difusão da poesia no ciberespaço. Computadores e técnicas de busca disponibilizadas na internet estendem a viabilidade de convergência de compatibilidades métricas, assim como em arquivos musicais e sonoros, que dividem com a poesia tal sinergia estética e vêm se destacando no cenário digital devido à sua popularidade.

1.2 PROBLEMAS E PREMISSAS

Os seres humanos têm uma grande capacidade de reconhecer e identificar a natureza literária de um texto. Entretanto, a criação de algoritmos precisos para a classificação de textos ainda é uma tarefa desafiadora para a sociedade atual (TIZHOOSH; SAHBA; DARA, 2008).

Com o crescimento das informações, principalmente no formato digital, a pesquisa na área de categorização de textos tem aumentado na última década, com o intuito de prover um acervo de conhecimentos particularizado ao público. Isto pode levar ao desenvolvimento de algoritmos que simulam o processamento humano, que, por sua vez, podem ser aplicados nesta determinada área para reconhecimento de padrões e extração de dados.

Data mining é um método de extração de dados que pode ser considerado como o resultado da evolução proveniente de informações pelo meio digital (HAN; PEI; KAMBER, 2009). A mineração de dados, como também conhecida, pode ser definida como o processo de descobrimento de correlações, padrões e tendências significativas através da busca em grandes quantidades de dados armazenados em repositórios (GARTNER, 2016).

Como parte da estrutura textual, a poesia representa uma base de dados grande o suficiente para uma possível caracterização de textos. Tal forma de linguagem é representada por elementos tais quais ritmo e métrica. O ritmo é o fluxo de som produzido pelo poema e métrica é o padrão de repetição do ritmo (TURCO, 1986).

De tal modo, a análise automatizada da rima poética deve ser capaz de identificar todas as estruturas de rima dentro do verso, baseando-se no processamento digital de sinais para a padronização rítmica de cada poema. Assim, como parte da mineração de dados, após

selecionar, limpar os dados, codificar os mesmos, será possível aplicar conhecimentos de *data mining* para uma possível correlação entre as características espectrais apresentadas em diversos poemas e seus respectivos autores.

Um resumo pragmático do processo de extração de conhecimento é apresentado na Figura 1. No resumo, a primeira etapa, de pré-processamento, corresponde à captação, organização e tratamento de dados. A segunda etapa compreende a mineração de dados, com a busca efetiva por conhecimentos úteis para a extração. A terceira etapa, de pós-tratamento, abrange o tratamento do conhecimento obtido na mineração de dados (GOLDSCHMIDT; PASSOS, 2005).

Figura 1. Etapas operacionais do processo de extração de conhecimento.

Fonte: Autoria Própria.

Assim, na fase de pré-processamento o intuito é criar uma base de dados constituída de poemas tradicionais da língua portuguesa, organizados de acordo com seus autores e escolas literárias. Nesta fase, também, haverá o tratamento destes dados, para que todos os poemas obedeçam a uma regra de padronização, com o processo de codificação de letras e símbolos gráficos em valores numéricos (com a utilização da tabela ASCII, por exemplo).

O texto codificado possibilita a análise da amplitude dos caracteres em função do período estabelecido por cada poema. Investiga-se, assim, a possível existência de uma estrutura poética padronizada de acordo com sua formação rítmica, evidenciada por seu comportamento periódico causado pela métrica disposta no texto repetitivo.

Por conseguinte, na mineração de dados, contemplando conhecimentos de processamento digital de sinais, haverá a análise do conteúdo extraído. O propósito é criar uma correspondência entre as rimas e métricas de cada poema com seus respectivos autores e, possivelmente, escolas literárias.

Por fim, no pós-processamento, haverá a efetiva classificação dos dados analisados e o paralelismo entre as características dos poemas poderá ser verificado.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo geral do presente trabalho baseia-se no estudo experimental de classificação de estruturas poéticas língua portuguesa de acordo com seus autores e, possivelmente, escolas literárias, utilizando técnicas de processamento digital de sinais e data mining.

1.3.2 Objetivos Específicos

- Codificar letras e símbolos gráficos em valores numéricos;
- Converter textos em sequências numéricas;
- Trabalhar com processamento digital de sinais e *data mining* para avaliar a cadência e identificação métrica de poemas;
- Analisar a correspondência de atributos referentes à codificação de sinais gráficos;
- Analisar dados provenientes da implementação para uma possível identificação autoral.

1.4 JUSTIFICATIVA

A sociedade atual tem capacidade de processar um grande número de informações de forma ativa, facilitando o processo de aprendizagem e difundindo, de modo crescente, o produto cultural decorrente do surgimento das formas de interação com as pessoas.

O hipertexto digital tem a premissa de automatizar e materializar operações de leitura, ampliando consideravelmente seu alcance. Assim, as bases de dados provenientes do conteúdo midiático reencontram a sensibilidade ao contexto das tecnologias somáticas (LEVY, 2007). Entretanto, não é de reconhecimento público a utilização de recursos capazes de automatizar textos provenientes da estrutura poética.

Atualmente, com a maior disseminação de informações no meio eletrônico, pesquisas na área de caracterização de dados estão em crescente desenvolvimento. Algoritmos para classificação de texto foram desenvolvidos para categorizar notícias, patentes, e-mails, artigos de revistas ou qualquer outra classificação para obter informações a partir da internet (TIZHOOSH; SAHBA; DARA, 2008). Entretanto, dada a inexistência de evidências teóricas sobre o assunto, os métodos criados para identificação textual não parecem se aplicar à poesia.

Nos últimos anos, a indústria da música, por exemplo, tem migrado cada vez mais para a distribuição digital através de lojas de música on-line e serviços de *streaming* como o Youtube, iTunes, Spotify, Grooveshark e Google Play (OORD; DIELEMAN; SCHRAUWEN, 2006).

Como resultado, a recomendação automática de músicas permite ouvintes a descobrir novas músicas que coincidem com os seus gostos. Esta recomendação tende a comparar músicas que são perceptivelmente semelhantes ao que o usuário tenha escutado anteriormente, através da medição da semelhança entre os sinais de áudio. Esta análise pode, então, de maneira similar, ser aplicada à poemas.

A justificativa deste trabalho apoia-se, assim, no fato de que é necessário aprofundar conhecimentos na área de *data mining* e processamento digital de sinais para fins de classificação de dados. Entretanto, ainda não está claro quais são as características importantes para o estudo deste sistema utilizando técnicas para identificação e classificação de poemas (TIZHOOSH; SAHBA; DARA, 2008).

De tal modo, é evidente a necessidade de estudo de uma nova ferramenta de reconhecimento de características provenientes de estruturas poéticas com o intuito de incluir este segmento no desenvolvimento de novas competências textuais para o meio digital.

1.5 PROCEDIMENTOS METODOLÓGICOS

O Trabalho de Conclusão de Curso apresentado fundamenta-se em uma metodologia qualitativa. Ressalta-se que uma avaliação qualitativa "[…] é caracterizada pela descrição, compreensão e interpretação de fatos e fenômenos, em contrapartida à avaliação quantitativa, denominada pesquisa quantitativa, onde predominam mensurações" (MARTINS, 2008).

Assim, com a inexistência de pesquisas exclusivamente voltadas ao presente tema, um estudo de caráter exploratório e investigativo serve como base para os procedimentos metodológicos desta pesquisa. O método de estudo é embasado em uma investigação empírica, onde o pesquisador não tem controle sobre eventos e variáveis, buscando apreender a totalidade de uma situação e, criativamente, descrever, compreender e interpretar a complexidade de um caso concreto (MARTINS, 2008).

Pretende-se, então, catalogar um determinado número de poemas tradicionais na língua portuguesa, previamente processando estes dados de acordo com uma padronização estética estipulada. Assim, utiliza-se desta base de dados para a conversão de sinais gráficos em números, sendo possível o processamento digital destes sinais. Através da análise matemática,

realizada no *software* MATLAB, é possível, de tal forma, categorizar os poemas selecionados de acordo com os parâmetros pré-estabelecidos, como autoria e escolas literárias dos poemas.

1.6 ESTRUTURA DO TRABALHO

O atual Trabalho de Conclusão de Curso será constituído de cinco capítulos. Inicialmente, o primeiro capítulo apresentará um panorama abrangente do tema, com a introdução, delimitação do tema, problemas e premissas, objetivo e justificativa do trabalho.

O segundo capítulo será constituído de uma pesquisa teórica que abrangerá uma revisão bibliográfica sobre os temas contemplados no trabalho, tais como a conceituação e definição do conteúdo essencial da literatura brasileira, a fim de entender e justificar sua correlação com o processamento digital de sinais e *data mining*.

O terceiro capítulo, baseado na conceituação teórica anteriormente definida em suma, compreenderá um conteúdo dissertativo e explicativo referente ao método de aplicação dos processos citados para a exploração de dados a fim de encontrar padrões relevantes na estrutura literária poética.

O quarto capítulo será composto pela apuração dos dados obtidos por meio da aplicação proposta no capítulo anterior, com o propósito de obter um resultado capaz de correlacionar poemas e autores através das densidades espectrais de energia encontradas nas estruturas literárias analisadas.

Por fim, o quinto, e último capítulo, concluirá este Trabalho de Conclusão de Curso com uma análise abrangente sobre o desenvolvimento da aplicação do processamento digital de sinais e *data mining* na categorização de poemas.

2 CONCEITOS DA LITERATURA

Apresenta-se, neste capítulo, uma base contextual sobre a composição de escritos literários na língua portuguesa. Contemplam-se conceitos e definições referentes à estrutura poética das diferentes escolas literárias brasileiras relevantes para o estudo neste trabalho.

Assim, o intuito deste capítulo é delimitar parâmetros de análise dentro da literatura brasileira, além de identificar estruturas textuais e características principais das escolas literárias cabíveis às possíveis padronizações. Tal análise será útil com a aplicação de métodos para identificação de sistemas.

2.1 DEFINIÇÃO

A literatura, do latim "*littera*", define-se como a "arte de compor escritos, em prosa ou em verso, de acordo com determinados princípios teóricos ou práticos" (MICHAELIS, 2017). Refere-se à literatura, então, todo e qualquer texto escrito que constitui valor artístico ou intelectual devido ao uso da linguagem através de estruturas linguísticas que diferem do uso comum.

Diferente da literatura lato sensu, a literatura stricto sensu representa conjuntos de textos que se constituem de características como a ficcionalidade, plurissignificação e originalidade (BERNARDI, 1999). Dentro deste conceito, a literatura se estabelece como uma arte única, tal qual emana a harmonia textual.

Assim, artisticamente, a definição de literatura associa-se à ideia de estética e à constituição de procedimentos estéticos em textos. De modo prático, esta arte se verifica em prosa ou verso, cada qual explorando um aspecto da arte, tratando de poemas, contos, novelas, romances, epopeias e peças teatrais (TRINGALI, 1994).

Informações referentes à literatura no contexto matemático são escassas. Deste modo, é interessante investigar e possivelmente caracterizar elementos distintos a esta arte tão singular. Para isso, são necessários o estudo e a análise de conceitos literários úteis na identificação de sistemas.

2.2 ARTE EM VERSO

O verso poético, ou poesia, contrário à prosa, fundamenta-se em uma composição textual formada por rimas ou versos livres, em que o autor expressa seus sentimentos, ideias e impressões (MICHAELIS, 2017).

Versos, por conseguinte, são definidos por palavras ou reunião de palavras que representam uma unidade rítmica de um poema (MICHAELIS, 2017). Os notáveis arranjos harmônicos de poemas são formados por estrofes que se caracterizam, basicamente, pela composição de um conjunto de versos.

A versificação caracteriza-se pelo ato ou a arte de reunir palavras em versos ou, em outras palavras, em uma forma de expressão socialmente ou culturalmente reconhecida, gerando arranjos padronizados no tempo (DEWYEY; FROG, 2009). Tal característica é de extrema importância para o presente estudo, já que se julga possível relacionar o conceito da escrita padronizada com uma determinada modelagem matemática.

Sistemas versificados são dinâmicos, ocasionando um relacionamento entre a formulação de novas expressões e o estabelecimento de tradições envolvidas pelo período da escrita. A formulação de novas expressões literárias é caracterizada por conceitos inerentes à escrita versificada. Ritmo, metro, rima, cadência e formas fixas são concepções poéticas que constituem o caráter repetitivo de tais estruturas.

A forma poética é uma escrita que compartilha elementos estruturais com textos de natureza da fala. O ritmo, como nosso coração, que pulsa intercalando batidas e pausas, é caracterizado, dentro da literatura, pela repetição de unidades melódicas dispostas em um texto.

O ritmo, elemento fundamental para a presente análise, pode ser caracterizado como qualquer movimento recorrente regular e simétrico, geralmente representado pela sucessão regulada de elementos fortes e fracos, ou de condições opostas ou diferentes. Na poesia, este significado geral de recorrência regular ou padrão no tempo é verificado pela rima. Cabe destacar a compatibilidade entre o conceito de ritmo, em estruturas literárias, e oscilação rítmica em um sinal.

De tal forma, a composição de unidades rítmicas é baseada no estudo da métrica textual, técnica definida dentro da arte poética como o "estudo dos versos em relação à medida" (MICHAELIS, 2017). Ou seja, a métrica se baseia na numeração das sílabas gramaticais a fim de formar um verso.

Dentro do conceito poético, o metro é a estrutura rítmica básica de um verso ou linhas em verso. Muitas formas tradicionais de verso prescrevem um medidor específico, ou um

determinado jogo de medidores alternados em uma ordem particular, dando a caracterização referente a cada estilo ou época literária. Adicionalmente a este conceito, o estudo e o uso real de metros e formas de versificação são conhecidos como prosódia, facilitando o entendimento das formas fixas.

Complementar aos conceitos de ritmo e metro, a cadência designa o fluxo linguístico que descreve a entonação da voz e sua inflexão modulada com a ascensão e queda de seu som. A cadência, então, define a unidade de tempo, o pulso e o ritmo da composição ou de partes dela.

Na poesia, a cadência descreve o passo rítmico da linguagem, restaurando a qualidade audível à poesia como uma arte falada. Assim, similar ao compasso, a cadência "pode ser registrada mecanicamente, por meio das pausas e cesuras, ao passo que o ritmo resulta da sucessão de sons vocabulares acusticamente agradáveis fluindo no tempo segundo um movimento contínuo" (MOISÉS, 1974).

Assim, um poema é uma forma de literatura versificada que usa as qualidades estéticas e rítmicas da linguagem para evocar significados além do significado prosaico ostensivo, ou em lugar dele. Portanto, a partir da versificação e do conteúdo de estruturas poéticas, é possível analisar a aplicação de algoritmos para a identificação de padrões particulares à autores ou escolas literárias dos quais fazem parte. Tal identificação baseia-se, então, no comportamento rítmico e harmônico característico de tal estrutura literária.

2.3 ESCOLAS LITERÁRIAS

O desenvolvimento de textos literários depende, direta ou indiretamente, das condições concretas em que o poeta se situa. Assim, o ritmo e a métrica, relevantes para a identificação de sistemas, devem ter uma relação de concordância com a época em que a estrutura é efetuada.

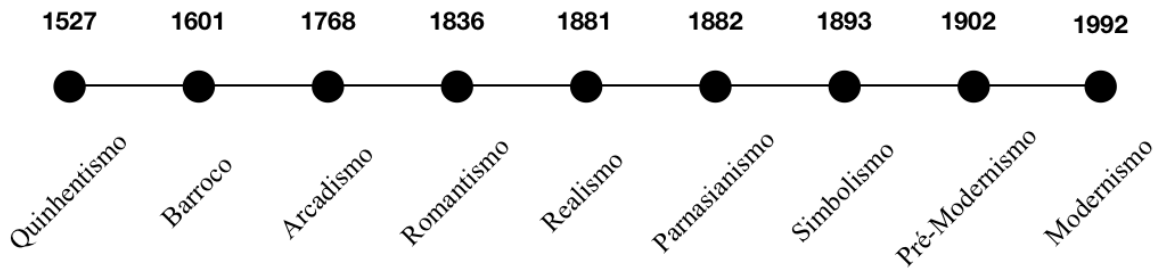
Caracterizado como um sistema coletivo de tendências, a palavra "escola", que do grego significa "ócio", tende, em âmbito artístico, construir uma filosofia de vida, uma estética e uma poética conforme a individualidade de sua época (TRINGALI, 1994).

As escolas literárias da língua portuguesa podem ser caracterizadas, também, como movimentos sociais, que têm como intuito a exploração de um diferente aspecto da arte conforme sua respectiva tratativa.

De tal modo, as escolas tais quais Quinhentismo, Barroco, Arcadismo, Romantismo, Realismo, Simbolismo, Parnasianismo, Pré-Modernismo e Modernismo apresentam composições estruturais particulares ao momento referenciado pelos movimentos, como

demonstrado na Figura 2. Tais composições podem contribuir para a identificação de sistemas, apresentando possíveis padrões através de suas diferentes caracterizações.

Figura 2. Escolas literárias na linha do tempo.



Fonte: Autoria Própria.

2.3.1 Quinhentismo

O Quinhentismo se inicia em 1500 com um nobre marco histórico social brasileiro, o descobrimento do Brasil. Este movimento é caracterizado por seu caráter informativo, acompanhando o processo de miscigenação e colonização portuguesa no Brasil, e dando início às manifestações literárias no perímetro brasileiro através da Carta de Pero Vaz de Caminha (MOISÉS, 1999).

Assim, a atividade literária representativa do Quinhentismo ignora os propósitos artísticos da literatura. De tal modo, prevalece em tal escola a intenção doutrinária e pedagógica sobre a estética (MOISÉS, 1999). De tal forma, por tal motivo, esta escola não é cabível ao estudo por não apresentar as características estruturais necessárias.

2.3.2 Barroco

"Áspero, rude, tosco, mal polido" (BERNARDI, 1999, p. 63), o Barroco foi a primeira manifestação literária artística no Brasil, tendo seu início em 1601. Tal expressão em forma de arte se apresenta como resposta aos reflexos dos ensinamentos conservadores provenientes da Renascença.

A elaboração de artifícios contraditórios dentro da estrutura literária barroca se contextualiza em uma constante preocupação com a estética. Esta harmoniosa estrutura

concentra-se na utilização de figuras de linguagem, demonstrando o intuito de maravilhar o leitor através da escrita (SILVA; SANT'ANA, 2007). Tal cuidado com formas literárias pode ser observado em obras de Gregório de Matos e Bento Teixeira.

2.3.3 Arcadismo

A ideologia burguesa, em oposição à demasia cultural do movimento barroco, emerge no cenário social brasileiro em 1768 como um propulsor de questões mundanas e simples (SILVA; SANT'ANA, 2007). O Arcadismo, assim, caracteriza-se pela adoção de gêneros e formas consagradas pela tradição (SILVA; SANT'ANA, 2007).

A transformação cultural contemplada no Arcadismo é resultante da valorização do pensamento racional da época. Filósofos da época, como Rousseau, partem da premissa de que "o homem nasce bom, mas a sociedade o corrompe, devendo, portanto, retornar para a natureza" (LIMA, 2009). Assim, a volta às regras de arte rígidas do Renascimento é justificada pela expressão racional da natureza, impulsionando atividades simplistas com o intuito de aproveitar o dia (*Carpe diem*).

2.3.4 Romantismo

Introduzida em 1836, distanciando dos conceitos arcadistas, o Romantismo rompe com os modelos clássicos e surge com atitudes e escritos referentes a um mundo idealizado (LIMA, 2009). A liberdade de expressão, característica do movimento, é conduzida por retratos emotivos que valorizam todas as coisas (BERNARDI, 1999).

Com o Romantismo, surge, então, uma poesia voltada para o individualismo (SILVA; SANT'ANA, 2007). Impera, neste movimento, o sentimento de insatisfação, depressão e melancolia em redor do incompreendido, estimulando uma temática específica brasileira (CASTELLO, 2004).

2.3.5 Realismo

Em repúdio à escola Romântica, o Realismo surge em 1865 (LIMA, 2009). Este movimento exprime, então, a necessidade de abordar o retrato fiel da vida, contrariando a artificialidade dos escritos românticos.

O realismo expandiu a necessidade de representar e preocupação social como marca da literatura da época (OLIVEIRA, 2008). Assim, a mentalidade realista concretiza a objetividade dentro da criação artística (SILVA; SANT'ANA, 2007).

2.3.6 Parnasianismo

A valorização exagerada da forma dá-se com o Parnasianismo em 1882 (LIMA, 2009). Poetas desta época buscavam a perfeição formal, apresentando temáticas baseadas em um texto contrário à emotividade. Buscam a objetividade valorizando o ideal da arte, com certo preciosismo ao ressaltar singularidades em detalhes (BERNARDI, 1999).

2.3.7 Simbolismo

O ano de 1893 marca o início do Simbolismo no Brasil (LIMA, 2009). Dentro desta escola, destaca-se a utilização de fatores comuns a outros movimentos literários brasileiros, com reações contrárias à ciência e disciplina social.

O símbolo é o recurso mais adequado para exprimir aspirações e impulsos humanos, tornando-se fundamental retratar a emoção, a sensibilidade e os instintos obscuros impondo a revolução formal com o verso libertado (CASTELLO, 2004). O Simbolismo, em sua maneira, utiliza características como sinestesia e musicalidade a fim de "libertar a criação artística e sua expressão correspondente" (CASTELLO, 2004, p. 331).

2.3.8 Pré-Modernismo

Em 1902, o Pré-Modernismo exprime o desejo de libertação aos conceitos já abordados no passado (LIMA, 2009). Trata-se, então, resumidamente, de um período de transição entre a estética simbolista e a modernista.

De tal forma, ainda que de caráter inovador, falta ao Pré-Modernismo nitidez de traços característicos (BASTOS, 2004). Assim, este movimento caracteriza-se por apresentar elementos de todos os estilos de épocas anteriores. O Pré-Modernismo, surge, então, como uma época de transformação dos valores estéticos literários.

2.3.9 Modernismo

A Semana de Arte Moderna, em 1922, foi responsável por revolucionar os conceitos artísticos previamente estabelecidos na literatura brasileira (BARBOSA; SANTOS, 2009). Busca-se quebrar os paradigmas da literatura tradicional, deixando de lado traços estruturais, mas priorizando o significado do texto.

O Modernismo, assim, surge com o intuito de propor uma revisão da cultura brasileira, tendo o verso livre como traço mais evidente na escrita (BASTOS, 2004). Neste movimento, o ritmo tornou-se independente da métrica, criando uma liberdade estrutural, mas, ainda assim, com um sistema de significados expressivos (BASTOS, 2004).

3 CONCEITOS DE *DATA MINING* E PROCESSAMENTO DE SINAIS

A crescente disponibilidade de dados em forma digital desenvolve o interesse pelo processamento de sinais. De tal modo, o intuito do presente capítulo é exprimir a correspondência entre a disponibilidade de dados, provenientes de qualquer estímulo, e o processamento de sinais.

Consequentemente, dentro deste capítulo serão apresentados conceitos e definições de conteúdos técnicos, a fim de prover subsídios para uma futura análise referente às estruturas literárias anteriormente apresentadas.

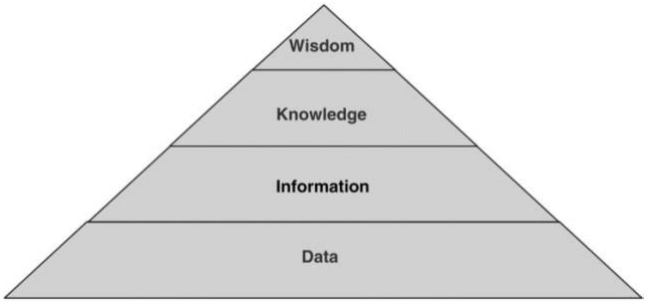
3.1 *DATA MINING*

A tecnologia da informação tem quebrado paradigmas comunicacionais de modo rápido e eficaz através da popularização de novos dispositivos e ferramentas, que têm como objetivo transmitir um maior número de informações à sociedade (HAN; PEI; KAMBER, 2009).

Esta grande disponibilidade e acúmulo de informações são características de uma inter-relação dinâmica com a liberdade humana. Isto quer dizer que o aumento do conhecimento adquirido determina a possibilidade de novas ações, que podem expandir a responsabilidade do indivíduo (AGAZZI; MINAZZI, 2008).

De tal modo, a relação essencial entre informação e sabedoria está contemplada dentro da hierarquia DIKW (*Data Information Knowledge Wisdom*), ou também conhecida como hierarquia do conhecimento. Esta hierarquia, proposta por Ackoff em 1989 (ACKOFF, 1989), estabelece os níveis de contribuição entre dados, informações, conhecimentos e sabedoria, como na Figura 3.

Figura 3. Hierarquia do conhecimento.



Fonte: ACKOFF, 1989.

Dados são definidos como símbolos, produtos da observação. Entretanto, são inúteis até que estejam em uma forma relevante. Inferida dos dados, obtém-se, então, a informação. Os sistemas de informação geram, armazenam, recuperam e processam dados. O conhecimento é o que torna possível a transformação de informação em instruções. Tal conceito pode ser obtido por transmissão, por instrução, ou extração da experiência. Por fim, a sabedoria é a capacidade de aumentar a eficácia da transmissão de conhecimento. A sabedoria agrega valor, o que requer a utilização de valores éticos únicos e pessoais (ROWLEY, 2007).

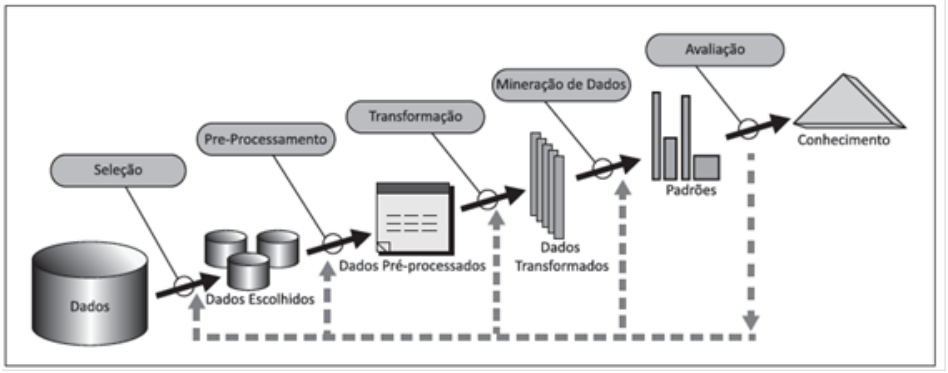
Assim, a capacidade de analisar dados e transformá-los em informações significativas são de extrema importância para o desenvolvimento da sociedade. Entretanto, uma maior disponibilidade de dados, com consequente aumento no tamanho e complexidade da base de informações, torna insustentável a gestão do conhecimento através, apenas, da análise de dados manual.

De tal forma, existe uma necessidade urgente de uma nova geração de teorias computacionais e ferramentas para auxiliar os seres humanos a extrair informações úteis (conhecimento) dos volumes de dados digitais em rápido crescimento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Tais teorias e ferramentas são contempladas dentro do campo de estudo emergente chamado *Knowledge Discovery in Databases* (Extração de Conhecimento em Bases de Dados).

A extração de conhecimento em bases de dados é um processo exploratório de análise e modelagem de grandes repositórios. Este processo é definido como um método não trivial para identificar padrões válidos, úteis e compreensíveis em dados (CIOS et. al, 2007).

O termo "processo" implica que a descoberta de conhecimento compreende muitas etapas, que envolvem preparação de dados, busca de padrões, avaliação de conhecimento e refinamento, como explícito na Figura 4.

Figura 4. Processo de extração de conhecimento.



Fonte: FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996.

Assim, a mineração de dados envolve a confecção de algoritmos para explorar dados, desenvolver modelos e descobrir padrões até então desconhecidos (MAIMON; ROKACH, 2005). De tal forma, o método de mineração de dados é utilizado em um subconjunto de dados para o entendimento, análise e predição de fenômenos obtidos em dados.

Portanto, a extração de um padrão também designa adequar um modelo aos dados, encontrar a estrutura a partir dos dados, ou, em geral, fazer uma descrição de um conjunto de dados. Assim, a obtenção de padrões em sistemas, essencial ao processo de descoberta de conhecimento, é dada, dentro na mineração de dados, de três formas: análise exploratória de dados, modelagem descritiva e modelagem preditiva (HAND; MANNILA; SMYTH, 2001).

As técnicas de análise exploratória de dados caracterizam-se por interpretações interativas e visuais. Ou seja, como parte inicial do processo de análise de dados, é importante sugerir hipóteses sobre as causas dos fenômenos observados, avaliar suposições sobre quais inferências estatísticas serão baseadas e fornecer uma base para a obtenção de dados através de inquéritos ou experiências (BEHRENS, 1997). Parte-se do princípio que tal técnica baseia-se em uma análise cética, mas ao mesmo tempo aberta, já que os dados devem ser sumarizados de forma presuntiva e, ao mesmo tempo, estar disposto a encontrar padrões inesperados como resultado da análise.

A modelagem descritiva, então, apresenta de uma forma conveniente as principais características dos dados. Esta análise tem como foco descrever e quantificar os eventos e as relações entre fatores analisados (OLSON, 2017). Assim, torna-se importante a utilização desta etapa no processo de mineração de dados, pois permite o entendimento do comportamento de fenômenos analisados que podem influenciar tratativas futuras.

As tratativas obtidas através da modelagem descritiva podem ser investigadas pela modelagem preditiva. Esta modelagem caracteriza-se pelo processo no qual um modelo é criado ou escolhido na tentativa de prever, com certa exatidão, possíveis comportamentos ou classificações cabíveis ao sistema (OLSON; WU, 2017). Ou seja, com a interpretação do comportamento dos dados analisados, pode-se prever resultados e padrões provenientes dos dados.

Assim, a disponibilidade de dados e, consequentemente, de informações, serve como base para a busca de conhecimento. A mineração de dados, então, permite a identificação e classificação de dados a fim de buscar e prever soluções para problemas propostos na sociedade.

3.2 PROCESSAMENTO DE SINAIS

O processamento de sinal é uma tecnologia que abrange a teoria fundamental, aplicações, algoritmos e implementações de processamento ou transferência de informações contidas em muitos formatos físicos, simbólicos ou abstratos diferentes, amplamente designados como sinais.

3.2.1 Processamento de Sinais

Toda e qualquer aplicação de processamento de sinais tem como base um modelo matemático de entrada e saída do sistema, que é analisado para prever o comportamento do sistema como um todo.

3.2.1.1 História

As especulações referentes à utilização do processamento digital de sinais têm suas raízes no período da Segunda Guerra Mundial. Nesta época, a aplicabilidade de técnicas digitais para análise de sinais foi considerada, inicialmente, para funções de filtro. Entretanto, com a inexistência de elementos de *hardware* necessários, o custo, o tamanho e a confiabilidade dos equipamentos já existentes fez com que a utilização de implementações analógicas permanecesse na vanguarda do processamento digital de sinais.

Com o crescente desenvolvimento de componentes, a implementação de filtros digitais se tornou cada vez mais econômica e viável (STRANNEBY, 2001). Até que em 1960 e 1970 maiores aplicações se tornaram possíveis devido à crescente disponibilidade de computadores.

De tal forma destaca-se o papel do processamento de sinais como uma ferramenta necessária para o desenvolvimento tecnológico, social e cultural. Tal ferramenta refere-se às várias técnicas para melhorar a precisão e confiabilidade das comunicações. Assim, conclui-se que estudos teóricos por trás do processamento de sinais abrangem procedimentos para esclarecer ou padronizar, também, os níveis ou estados de um sinal digital, conforme avanços tecnológicos.

3.2.1.2 Conceito

O conceito de processamento de sinais é fundamentado na representação do sistema de modo otimizado, a fim de extrair informações baseando-se em uma fonte de informação, conforme exemplificado na Figura 5.

Figura 5. Modelagem para o processamento de sinais.

Fonte: RABINER, 1978.

Assim, o processamento de sinais refere-se a qualquer operação que modifica, analisa ou manipula, de algum modo, a informação contida em um sinal (PRANDONI; VETTERLI, 2008).

Adicional ao conceito do processamento de sinais, o adjetivo “digital” deriva de *digitus*, do latim “dedo”. Esta palavra descreve consistentemente uma visão ampla, onde tudo pode ser representado por um número inteiro (PRANDONI; VETTERLI, 2008). Consequentemente, tal representação binária indica que o tempo e a amplitude de um sinal são quantidades discretas.

De tal modo, a implementação do processamento digital “baseia-se em transformar o sinal analógico em sinal discreto, transformar o sinal discreto em um sinal digital, processar o sinal digital, transformar o sinal digital processado em um sinal discreto e transformar o sinal discreto processado em um sinal analógico” (DE LA VEGA, 2017).

3.2.1.3 Método

A necessidade de discretizar os valores de amplitude de um sinal de tempo discreto vem do fato de que, no mundo digital, todas as variáveis são necessariamente representadas com uma precisão finita. Assim, a conversão do valor analógico de um sinal para sua contraparte digital discretizada é chamada conversão analógico-digital (PRANDONI, VETTERLI; 2008).

A amostragem, caracterizada pela conversão do sinal analógico para o sinal discreto, relaciona, então, a velocidade na qual é preciso medir repetidamente o sinal com a frequência máxima contida em seu espectro. Tais espectros são calculados usando a transformada de Fourier.

Originalmente proposta em 1807 pelo matemático francês Jean Baptiste Joseph Fourier (STRANNEBY, 2001), o método da transformada é utilizado para reduzir a complexidade de operações através da mudança do domínio dos operadores a fim de tornar um sinal digital.

Assim, um sinal analógico pode ser descrito em relação à sua frequência através do cálculo da transformada de Fourier. Na equação, x é a variável independente do sistema, f(x) é a função independente do sistema e ω representa a frequência angular envolvida na variação de f(x) (MORITA, 1995).

$$F(\omega) = \int_{-\infty}^{\infty} f(x) e^{-jx\omega} dx$$

Consequentemente, a operação que recupera os dados da sequência discreta da função é chamada transformada inversa de Fourier (MORITA, 1995).

$$f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(\omega) e^{-jx\omega} d\omega$$

Ainda que a transformada e a transformada inversa de Fourier demonstrem como o domínio da frequência se difere das amplitudes senoidais do sinal, é definido o conceito de densidade espectral para entender como a energia do sinal está distribuída na frequência.

O espectro é uma representação completa e alternativa de um sinal. Através da análise do espectro, pode-se obter, de relance, a informação fundamental, necessária para caracterizar e classificar um sinal no domínio da frequência. Assim, o espectro de energia de um sinal é dado pela seguinte equação (VASEGHI, 2008).

$$E(f) = |F(\omega)|^2$$

Conclui-se, então, que com a análise matemática do sinal através do processamento digital, adquirem-se informações relevantes referentes a sequências periódicas. Observa-se uma correspondência entre a frequência do sinal e, consequentemente, seu período, com seu

significado analógico, verificando possíveis simetrias na estrutura poética analisada devido seu comportamento harmônico.

4 EXTRAÇÃO DE CONHECIMENTO EM ESTRUTURAS POÉTICAS

A necessidade de ampliar as capacidades de análise humana para lidar com o grande número de informação que podemos coletar é de extrema importância científica. Os dados que capturamos sobre nosso ambiente são a evidência básica que usamos para construir teorias e modelos do universo em que vivemos.

O reconhecimento de padrões constitui a base de cada ciência natural: as leis da física, a descrição das espécies na biologia ou a análise do comportamento humano; todos baseados em identificar padrões. Também na vida diária, o reconhecimento de padrões desempenha um papel importante, neste caso ao ler textos, identificar autores e descobrir padrões em poemas.

Como os computadores permitiram que os humanos reunissem mais dados do que podemos digerir, é natural recorrer a técnicas computacionais para nos ajudar a descobrir padrões e estruturas significativas em dados. De tal modo, o conceito da extração de conhecimento em estruturas poéticas é aplicado conforme a Figura 5.

4.1 SELEÇÃO DE DADOS

O intuito principal da seleção de dados é determinar tipos e fontes de dados adequadas para a coleta dos mesmos. Desta forma, objetiva-se estruturar, dentro do contexto específico, elementos que tenham relevância significante quando comparados a todo universo da literatura poética brasileira, de forma a eliminar dados que se apresentem dispensáveis ao cenário.

Assim, os escritos artísticos que compõem a literatura brasileira, tanto em forma de poesia quanto em verso, são selecionados inicialmente com a aquisição apenas, e somente apenas, de estruturas literárias poéticas. Por conseguinte, através da vasta gama de dados disponíveis nesta forma, é relevante a apresentação de elementos que contemplem todas as escolas literárias analisadas neste estudo (Figura 4).

Com isto, a coleta de poemas, cabíveis às escolas literárias referenciadas, é feita de forma amostral. A constituição da amostra, neste caso, é baseada em autores e poemas que apresentem maior notoriedade dentro do âmbito poético. Desta forma, uma amostragem inicial de 184 poemas é coletada a fim de garantir a representatividade na busca por padrões em estruturas poéticas, como explicitado na Tabela 1.

Tabela 1. Autores, escolas e quantidade de poemas selecionados como base de dados.

Escola Literária	Quantidade	Autor
Arcadismo	31	Basílio da Gama Cláudio Manuel da Costa Tomás Antônio Gonzaga
Barroco	11	Bento Teixeira Gregório de Matos
Modernismo	20	Manuel Bandeira Oswald de Andrade
Parnasianismo	30	Alberto de Oliveira Olavo Bilac Raimundo Correia
Pré-Modernismo	10	Lima Barreto
Realismo	20	Machado de Assis Mario de Andrade
Romantismo	31	Alvares de Azevedo Castro Alves Fernando Pessoa Gonçalves Dias
Simbolismo	30	Alphonsus de Guimarães Augusto dos Anjos Cruz e Sousa

4.2 PRÉ-PROCESSAMENTO

A etapa de pré-processamento dos dados provenientes de estruturas poéticas da língua portuguesa compreende a formatação dos dados a fim de posteriormente utilizá-los em algoritmos para a mineração de dados. Isto é, a partir dos poemas originais analisados, obtém-se derivações estruturais, adicionando, modificando ou removendo atributos inicialmente estabelecidos.

No presente trabalho, a estética, elemento crucial do gênero textual poético, é analisada considerando a transformação inicial de caracteres constituintes da estrutura. Ou seja, poemas, compostos por letras e sinais gráficos, têm seus caracteres investigados a fim de prepará-los para futuras interpretações.

De tal forma, a investigação inicial consiste em formatar a estrutura textual a fim de verificar a representatividade dos símbolos constituintes no poema. Assim, a influência e relevância dos caracteres foram utilizadas através das seguintes formas:

- Poema original, com nível nulo de modificações, composto por letras maiúsculas e minúsculas, sinais gráficos e acentuação;
- Poema sem letras maiúsculas, levando em consideração apenas letras minúsculas no texto, substituindo toda e qualquer letra maiúscula por sua equivalente minúscula;
- Poema sem sinais de pontuação, com a exclusão de todos os sinais gráficos da estrutura, mas com a acentuação originalmente proposta no poema;
- Poema sem acentuação, levando em consideração apenas letras não acentuadas no texto, substituindo toda e qualquer letra acentuada por sua equivalente;
- Poema sem sinais "invisíveis", tais como espaço e quebra de linha, mantendo as palavras em forma contínua;
- Poema sem consoantes ou vogais, eliminando da estrutura o caractere de classe oposta;
- Poema sem os primeiros elementos de cada frase, mantendo somente os últimos 25% de cada verso, a fim de identificar uma possível cadência entre versos;

O comportamento da estrutura é estudado a fim de investigar a atuação de elementos expressivos no poema. A inclusão e modificação de elementos na estrutura permite uma maior adequação dos textos literários à investigação proposta, no sentido de que proporciona uma representação visual do comportamento do poema.

É importante frisar, neste momento, que a exclusão de sinais gráficos e acentuação em estruturas poéticas foram desvencilhadas da imagem acústica do poema. O referencial sonoro, constituído pelo fonema do segmento, não deve ser confundido com o referencial gráfico, formado pela estrutura escrita.

4.3 TRANSFORMAÇÃO

A transformação de dados consiste no simples processo de converter dados de um formato para outro. Intenciona-se, assim, consolidar a informação proveniente do pré-processamento em formas adequadas para a mineração de dados. De tal forma, as estruturas poéticas estudadas devem ser convertidas de acordo com padrões de codificação, a fim de representar caracteres alfanuméricos.

O uso de padrões para representação de caracteres aumenta a eficiência e elimina erros na mineração de dados. Devido sua significante utilidade em grandes grupos de usuários, a internet, por exemplo, configura a plataforma de comunicação de forma global, e, portanto, usar padrões é uma maneira de alcançar o objetivo de unificar a comunicabilidade.

De tal modo, a fim de codificar estruturas textuais, considera-se um caractere como uma unidade mínima de texto que tem valor semântico. Assim, consequentemente, um texto é constituído por uma coleção de caracteres. Portanto, um conjunto de caracteres codificados é um conjunto em que cada caractere corresponde a um número exclusivo.

Destaca-se a utilização de um elemento numérico para cada caractere do texto. Utiliza-se a análise unitária já que esta representa a menor unidade estrutural no texto, entretanto a mesma análise pode ser feita através de fonemas.

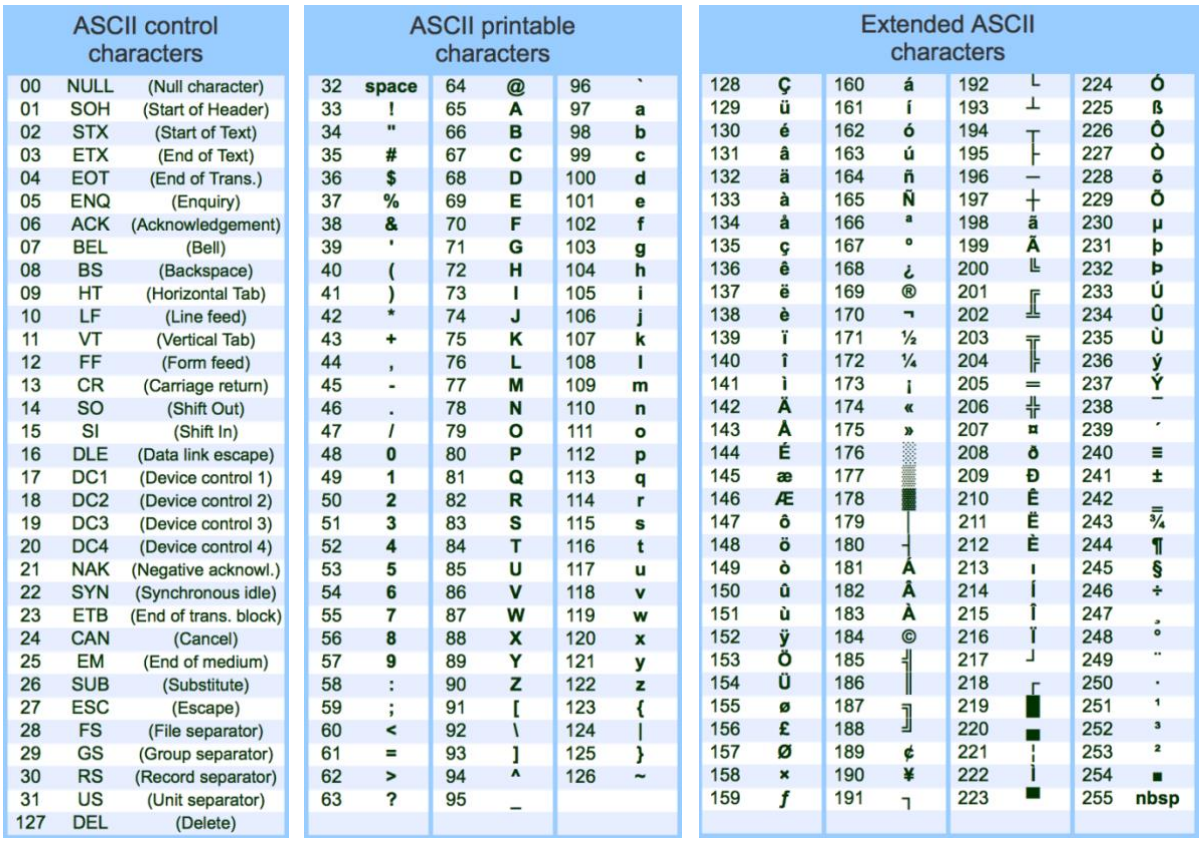
4.3.1 Transformação UTF-8 para ASCII

Neste contexto, como principal base para codificação das estruturas literárias analisadas no estudo, destaca-se a utilização da codificação UTF-8, *Unicode Transformation Format – 8 bit*. O Unicode é o padrão universal de codificação de caracteres para caracteres e texto escritos, definindo uma forma consistente de codificação de texto multilíngue que permite a troca de dados de texto internacionalmente (THE UNICODE CONSORTIUM, 2017).

Sendo assim, este formulário de codificação, de tamanho variável, preserva a representante codificada fazendo uso de unidades de código de 8 bits. De tal modo, devido à facilidade da utilização da codificação UTF-8 com sistemas existentes com base em ASCII, a transformação dos poemas selecionados ocorre com a utilização desta codificação.

A codificação ASCII, *American Standard Code for Information Interchange*, foi criada em 1963 pelo *American National Standards Institute* (ANSI), e compreende a codificação de 255 caracteres diferentes, como exibido na Tabela 2.

Tabela 2. Tabela de códigos ASCII.



ASCII control characters		
00	NULL	(Null character)
01	SOH	(Start of Header)
02	STX	(Start of Text)
03	ETX	(End of Text)
04	EOT	(End of Trans.)
05	ENQ	(Enquiry)
06	ACK	(Acknowledgement)
07	BEL	(Bell)
08	BS	(Backspace)
09	HT	(Horizontal Tab)
10	LF	(Line feed)
11	VT	(Vertical Tab)
12	FF	(Form feed)
13	CR	(Carriage return)
14	SO	(Shift Out)
15	SI	(Shift In)
16	DLE	(Data link escape)
17	DC1	(Device control 1)
18	DC2	(Device control 2)
19	DC3	(Device control 3)
20	DC4	(Device control 4)
21	NAK	(Negative acknowl.)
22	SYN	(Synchronous idle)
23	ETB	(End of trans. block)
24	CAN	(Cancel)
25	EM	(End of medium)
26	SUB	(Substitute)
27	ESC	(Escape)
28	FS	(File separator)
29	GS	(Group separator)
30	RS	(Record separator)
31	US	(Unit separator)
127	DEL	(Delete)

ASCII printable characters					
32	space	64	@	96	`
33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
36	$	68	D	100	d
37	%	69	E	101	e
38	&	70	F	102	f
39	'	71	G	103	g
40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46	.	78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r
51	3	83	S	115	s
52	4	84	T	116	t
53	5	85	U	117	u
54	6	86	V	118	v
55	7	87	W	119	w
56	8	88	X	120	x
57	9	89	Y	121	y
58	:	90	Z	122	z
59	;	91	[	123	{
60	<	92	\	124	|
61	=	93	]	125	}
62	>	94	^	126	~
63	?	95	_		

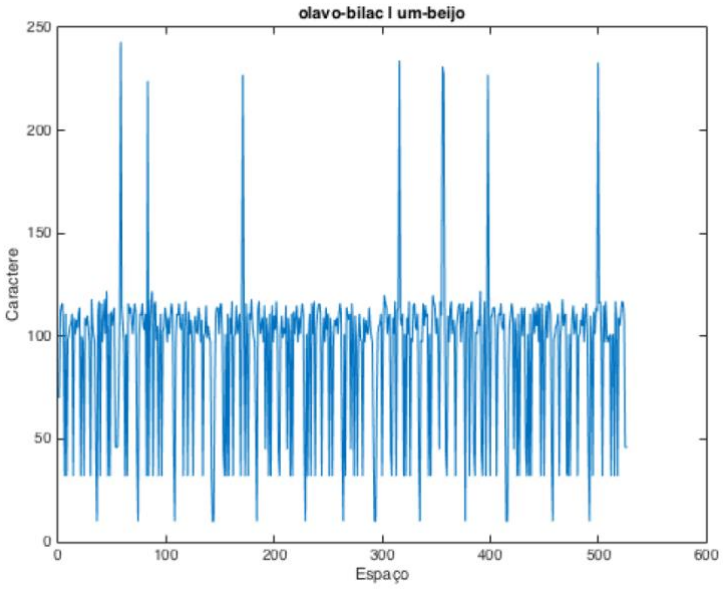
Extended ASCII characters							
128	Ç	160	á	192	└	224	Ó
129	ü	161	í	193	┴	225	ß
130	é	162	ó	194	┬	226	Ô
131	â	163	ú	195	├	227	Ò
132	ä	164	ñ	196	─	228	õ
133	à	165	Ñ	197	┼	229	Õ
134	å	166	ª	198	ã	230	µ
135	ç	167	º	199	Ã	231	þ
136	ê	168	¿	200	╚	232	Þ
137	ë	169	®	201	╔	233	Ú
138	è	170	¬	202	╩	234	Û
139	ï	171	½	203	╦	235	Ù
140	î	172	¼	204	╠	236	ý
141	ì	173	¡	205	═	237	Ý
142	Ä	174	«	206	╬	238	¯
143	Å	175	»	207	¤	239	´
144	É	176	░	208	ð	240	≡
145	æ	177	▒	209	Ð	241	±
146	Æ	178	▓	210	Ê	242	‗
147	ô	179	│	211	Ë	243	¾
148	ö	180	┤	212	È	244	¶
149	ò	181	Á	213	ı	245	§
150	û	182	Â	214	Í	246	÷
151	ù	183	À	215	Î	247	¸
152	ÿ	184	©	216	Ï	248	°
153	Ö	185	╣	217	┘	249	¨
154	Ü	186	║	218	┌	250	·
155	ø	187	╗	219	█	251	¹
156	£	188	╝	220	▄	252	³
157	Ø	189	¢	221	¦	253	²
158	×	190	¥	222	Ì	254	■
159	ƒ	191	┐	223	▀	255	nbsp

Fonte: http://www.theasciicode.com.ar

Como resultado da codificação de estruturas poéticas obtém-se a representação do texto poético de forma numérica. De tal forma, é possível descrever a comportamento do texto no espaço, a fim de buscar padronizações nas estruturas codificadas. Ou seja, a codificação de caracteres reflete a forma como o conjunto de caracteres codificados é mapeado para manipulação.

A Figura 6 abaixo mostra como caracteres e pontos de código no poema “Um Beijo”, de Olavo Bilac, em sua forma original, são mapeados para sequências usando a codificação ASCII.

Figura 6. Poema codificado em formato ASCII.



Fonte: Autoria Própria.

De tal forma, a partir da definição referente à codificação inicialmente utilizada no estudo, as diferentes modificações consideradas no pré-processamento podem, agora, ser transformadas e representadas de maneira gráfica. O intuito é apresentar graficamente modos distintos de tratar o sinal proveniente de estruturas poéticas, investigando a relevância de letras e sinais gráficos, como explicitado na etapa de pré-processamento.

4.3.2 Transformação por frequência de caracteres

Esta forma de análise objetiva recodificar estruturas literárias neste trabalho, mas requer um maior nível de abstração. Os elementos alfanuméricos constituintes de estruturas poéticas são dispostos nos textos de acordo com determinada frequência. Desta maneira, medir a quantidade de vezes que um mesmo caractere se repete na estrutura permite representar a similaridade ortográfica de palavras no texto.

Deste modo, os textos poéticos estudados são analisados conforme a disposição de seus caracteres na estrutura, de modo a identificar, contabilizar e codificar seus elementos de acordo com sua frequência. A partir da frequência com que os caracteres constituintes das estruturas poéticas aparecem no texto, é também possível suprimir aqueles que se comportam fora de uma determinada faixa de frequência. Através desta análise é possível relacionar a resposta comportamental da estrutura poética quando uma faixa de valores é rejeitada.

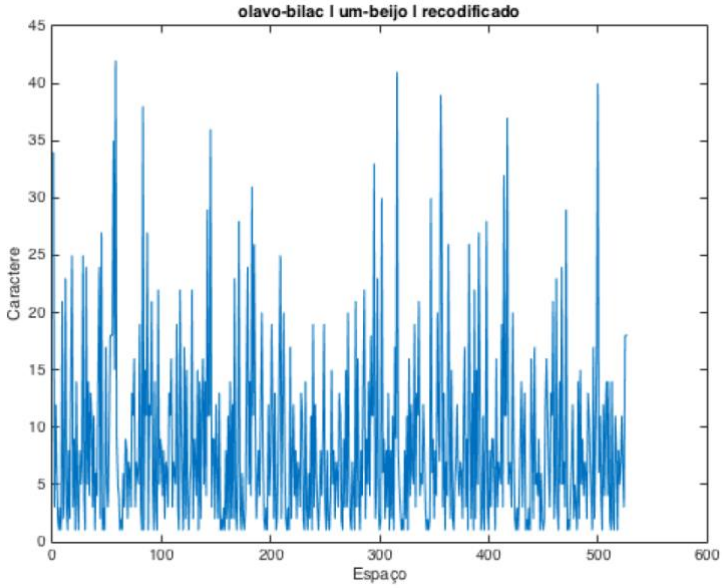
Como a presente análise tem foco nos grupos distintos de poemas (escolas literárias), a frequência de elementos é feita a partir da escola estudada. De tal forma, procura-se prever o comportamento do poema dentro do seu grupo de similaridade. A Tabela 3 mostra, por exemplo, a frequência de caracteres do Parnasianismo, com seus poemas em formato original.

Tabela 3. Frequência de caracteres no Parnasianismo.

Caractere	Frequência	Código	Caractere	Frequência	Código
[Espaço]	3094	1	S	41	39
a	2064	2	j	40	40
e	1871	3	N	37	41
o	1689	4	V	37	42
s	1236	5	P	35	43
r	1055	6	T	35	44
i	812	7	ê	34	45
n	796	8	x	31	46
m	754	9	O	29	47
d	746	10	ó	29	48
t	693	11	à	26	49
[Quebra de linha]	690	12	M	22	50
u	688	13	B	20	51
,	538	14	R	17	52
l	530	15	?	16	53
c	358	16	:	15	54
p	284	17	õ	15	55
v	265	18	ú	15	56
.	258	19	F	14	57
b	179	20	L	11	58
g	177	21	“	10	59
f	171	22	”	10	60
q	149	23	H	7	61
h	141	24	I	7	62
E	117	25	G	6	63
z	100	26	U	6	64
!	94	27	‘	5	65
-	80	28	É	5	66
ã	76	29	Ó	5	67
A	67	30	â	5	68
D	60	31	À	3	69
á	57	32	ô	3	70
é	54	33	(	1	71
ç	51	34	)	1	72
C	44	35	J	1	73
í	43	36	Í	1	74
;	42	37	Ú	1	75
Q	42	38			

Assim, como conseguinte, a Figura 7 abaixo mostra como caracteres do mesmo poema "Um Beijo", de Olavo Bilac, em sua forma original, são recodificados de acordo com a frequência que aparecem no texto.

Figura 7. Poema recodificado de acordo com frequência de caracteres.



Fonte: Autoria Própria.

Assim, o padrão posicional de letras em um texto é, consequentemente, o reflexo do número (e frequência) de outras palavras que compartilham estas mesmas letras identificadas em posições particulares.

O tratamento dos dados se dá de forma seletiva objetivando analisar o comportamento de elementos textuais, de forma individual, como vogais, consoantes, sinais de pontuação e acentuação e caracteres maiúsculos ou minúsculos. Como anteriormente estabelecido, todos os poemas contidos somente por letras minúsculas, sem caracteres acentuados e sem pontuação são considerados em sua forma "limpa".

De tal forma, todos os caracteres são recodificados baseados em sua ordem de frequência. Tal análise é aplicada na população total de poemas que, consequentemente, podem indicar a correlação entre estruturas de mesmo movimento literário ou autor.

Como resultante, a utilização de oito caracteres específicos (<espaço>, a, e, o, s, r, i, n) é predominantemente visível na distribuição probabilística de todas as escolas literárias. Este indício pode representar a relevância da utilização de tais caracteres no texto quando referente ao comportamento repetitivo de tais caracteres.

Desta forma, investiga-se o comportamento de caracteres conforme a estruturação imposta pelo movimento. A maior frequência relativa indica um maior coeficiente entre a frequência absoluta de dados e o número de elementos da amostra em questão. Ou seja, talvez seja possível analisar a utilização de um determinado elemento na estrutura.

Entretanto, destaca-se que nenhuma distribuição de frequência de letras é exata a um determinado idioma, uma vez que todos os escritores escrevem de forma ligeiramente diferente. No entanto, a maioria das línguas tem uma distribuição característica que é fortemente aparente em textos mais longos, sendo fácil a identificação da utilização de letras específicas de acordo com a língua.

Assim, é relevante na análise, também, a consideração de estruturas poéticas escritas em línguas estrangeiras. O objetivo desta análise é caracterizar e identificar a forma como um idioma pode alterar a estrutura textual, apresentando elementos peculiares às línguas.

4.4 MINERAÇÃO DE DADOS

Na presente pesquisa, aspectos referentes à forma de estruturas poéticas são tratados. A partir de uma base de dados constituída por poemas na língua portuguesa, testes e análises são realizados a fim de identificar possíveis classificações em tais estruturas. São estudadas novas formas de descrever poemas a fim de buscar a relação entre seus caracteres e suas disposições no texto poético.

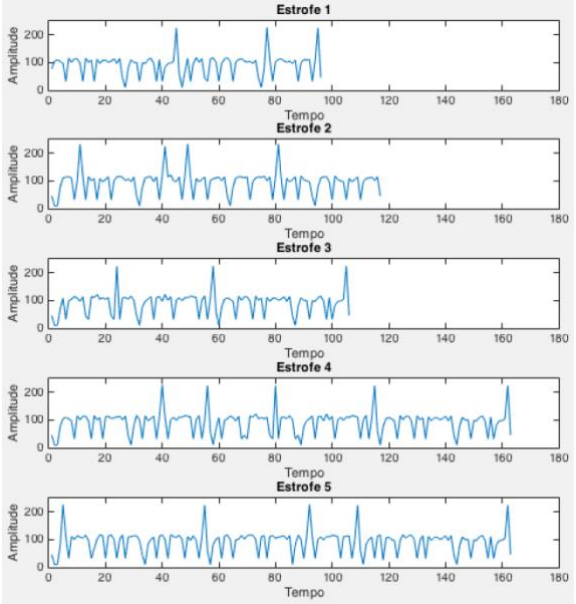
Através das análises propostas dentro do estudo de estruturas literárias, destaca-se a extração de conhecimento por meio da investigação e relevância de elementos nos poemas. De tal forma, as estruturas poéticas são estudadas após o processo de modificação de caracteres e adaptação de elementos em suas composições. Inicialmente, tal análise serve como base para síntese do conteúdo disposto, demonstrando pontos de coerência para a busca de padrões em textos poéticos.

A busca por padrões em textos poéticos abrange, então, a análise de todos os segmentos que compõe o texto. Desta forma uma visão geral referente à poemas é demasiadamente rasa, já que não apresenta, de forma visualmente clara, características notavelmente intrínsecas às estrofes e versos do poema.

Estudar o comportamento de estrofes propicia a investigação comparativa da composição poética. Ou seja, é necessária a análise de tal arranjo a fim de possivelmente identificar características elementares do poema. As estrofes devem conter, de modo geral, um comportamento similar, já que apresentam rimas que se harmonizam.

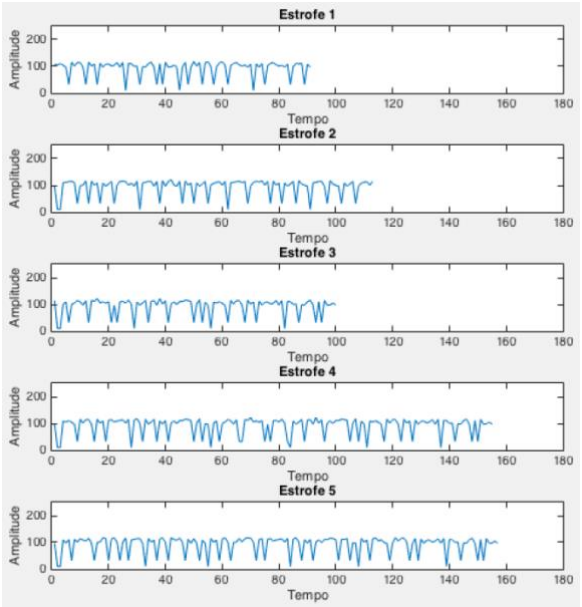
De tal forma, a fim de demonstrar o comportamento do poema “Canção do Exílio”, de Gonçalves Dias, são descritas as resultantes gráficas, na Figura 8 e Figura 9, da estrutura original (nível de modificação nulo) e da estrutura que considera a substituição e exclusão de todos os símbolos gráficos (nível de modificação máximo), respectivamente.

Figura 8. Estrofes de poema com nível de modificação nulo.



Fonte: Autoria Própria.

Figura 9. Estrofes de poema com nível de modificação máximo.



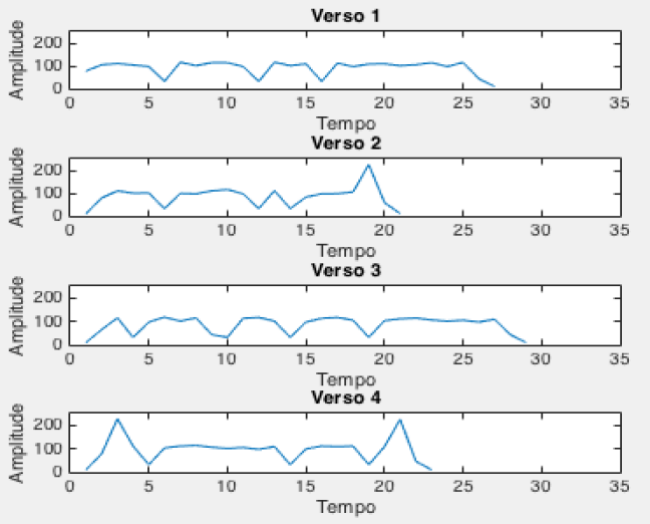
Fonte: Autoria Própria.

Portanto, características notavelmente intrínsecas aos poemas podem ser consideradas e identificadas através da análise de versos do poema. Versos são caracterizados como as

composições mais elementares de estruturas poéticas. Sendo assim, a busca por conhecimento e reconhecimento de padrões em poemas é diretamente dependente da análise comportamental de versos.

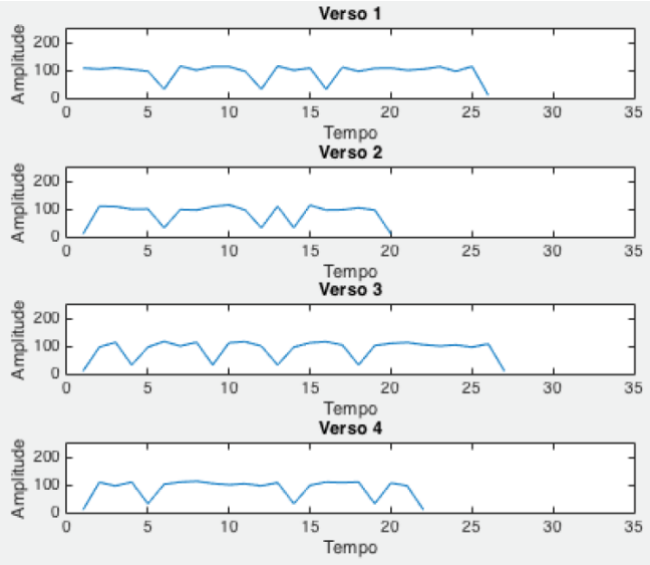
Desta forma, a Figura 10 e a Figura 11 apresentam, respectivamente, os versos contemplados na estrutura original (nível de modificação nulo) e na estrutura que considera a substituição e exclusão de todos os símbolos gráficos (nível de modificação máximo). O propósito é identificar ou prever, com maior precisão, as classificações possíveis do sistema.

Figura 10. Versos do poema com nível de modificação nulo.



Fonte: Autoria Própria.

Figura 11. Versos poema com nível de modificação máximo.



Fonte: Autoria Própria.

Portanto, com a obtenção de diferentes modelagens aplicadas aos textos poéticos, é possível notar a invariabilidade da estrutura poética quando colocada sob diferentes

modificações. Ainda que de forma exemplificada, no poema analisado é visível a similaridade entre o primeiro e o penúltimo verso e entre o segundo e o último verso da primeira estrofe. Obviamente, esta similaridade, obtida através da análise visual do texto, refere-se à uma rima. Prevê-se, assim, um padrão.

Desta forma, constata-se que, indiferentemente do nível de modificação com que a estrutura é adequada ou pré-processada, a resultante descritiva da estrutura é equivalente. Ou seja, as tratativas obtidas através da modelagem descritiva podem, de fato, induzir a modelagem preditiva. É possível, então, ainda que de forma inicial, prever padrões provenientes de versos obtidos em estruturas poéticas.

Portanto, a previsão de padrões em estruturas poéticas pode ser feita através da detecção do período fundamental de tais estruturas. De tal maneira, a classificação e identificação de elementos periódicos em textos poéticos dependem, dentro da mineração de dados, da similaridade comportamental do sistema, já apresentada.

4.5 AVALIAÇÃO

De maneira avaliativa, com esta fase fundamental do estudo de reconhecimento de padrões em estruturas poéticas, é possível iniciar a descoberta de conhecimentos disponíveis em tal formato textual. Desta maneira, diferentes técnicas de *data mining* são aplicadas objetivando demonstrar que partes destas estruturas, que são dependentes entre si, nos permitem detectar relações textuais complexas a um nível contextual elevado.

A detecção do período fundamental é interessante sempre que um sinal aparentemente periódico é estudado ou modelado, especialmente na fala e na música. Como consequência, a frequência fundamental define-se como a componente de menor frequência em uma onda senoidal complexa (OLSON, 1967). De tal modo, a detecção desta frequência permite que a mesma seja comparada com o período completo do sistema, detectando qualquer comportamento que seja anômalo à faixa de frequência típica do conjunto.

Tal análise pode ser agrupada em métodos no domínio do tempo e da frequência. No domínio da frequência o espectro do sinal periódico é analisado em busca de um comportamento harmônico obtido através da utilização de técnicas para análise temporal. Técnicas no domínio do tempo objetivam indicar picos provenientes da correlação cruzada ou autocorrelação do sinal.

A correlação cruzada é uma medida de similaridade entre dois sinais em função do deslocamento de um relativo ao outro. A autocorrelação pode ser considerada, assim, a

correlação cruzada de um sinal com ele mesmo. Portanto, se um sinal for periódico, ele será perfeitamente correlacionado com sua versão deslocada se o tempo de atraso for um número inteiro de períodos.

No entanto, é difícil construir modelos estatísticos confiáveis envolvendo o período fundamental devido a erros de estimativa de passo e a descontinuidade do espaço da frequência. Especialmente no caso de estruturas poéticas, é desconhecida a aplicação de tal metodologia avaliativa.

5 MÉTODO, DESENVOLVIMENTO E RESULTADOS

A análise matemática para identificação de padrões em estruturas literárias parte de um algoritmo criado com a utilização do *software* MATLAB. O intuito do programa desenvolvido é tratar e analisar os dados provenientes dos poemas selecionados de acordo com as modificações estruturais avaliadas durante a etapa de pré-processamento.

5.1 CONSIDERAÇÕES

O programa de computador desenvolvido é configurado para receber *inputs* do usuário a fim de selecionar, inicialmente, a escola literária e/ou autor a serem analisados e de que forma os poemas contemplados por esta escola e/ou autor serão tratados para posterior análise.

Com estas informações inicias, todos os poemas são concatenados em um vetor único, possibilitando o cálculo de frequência relativa de caracteres. De tal maneira, quando cada poema é individualmente lido, é realizado o cálculo de frequência dos caracteres deste poema quando comparados com todos os caracteres dos poemas dentro da mesma escola. Assim, é possível executar a transformação de caracteres tomando como base todos os elementos compostos na gama completa de poemas e sua frequência total.

A frequência dos caracteres é então normalizada. A normalização dos valores surge com o intuito de harmonizar escalas, correspondendo à adaptação da escala quando números muito diferentes estão em mesma dimensão. Desta forma, o valor de frequência máximo se equivale a 1, assim como o valor mínimo de frequência se equivale a 0. A normalização pode então ser acentuada quando elevada à um índice maior que 1. Assim, este passo resume-se na seguinte fórmula:

$$F_{eq} = \frac{F_{poemas}}{F_{máxima}}^{Índice}$$

Onde:

F_{eq} refere à frequência equivalente obtida com a normalização dos valores;

F_{poema} refere à frequência dos caracteres de todos os poemas da escola em questão;

$F_{máxima}$ refere à frequência máxima de caracteres obtida no poemas;

$Índice$ refere ao coeficiente de normalização.

Após a normalização do elemento poético, a recodificação de cada poema é realizada, como previamente citado na etapa de pré-processamento. Assim, a remoção do nível DC é feita com o intuito de remover valores de baixa frequência no sistema e que, consequentemente, não devem ter grande significância na análise.

Com o poema devidamente manipulado, é realizada a autocorrelação da estrutura, buscando provar matematicamente indícios de um sinal periódico contido nesta estrutura literária.

5.2 AUTOCORRELAÇÃO

A autocorrelação, elemento fundamental nesta análise, caracteriza-se como um caso particular da correlação cruzada. Trata-se do produto escalar de um sinal com ele mesmo quando deslocado no tempo. Assim, é possível medir o grau de similaridade (assim como de irregularidade) de um sinal com ele mesmo (GAYDECKI, 2004). Desta forma, é possível fazer esta análise através da seguinte fórmula:

$$y[n] = \sum_{k=0}^{M-1} x[k].x[n+k]$$

Onde:

$x[n]$ é o sinal a ser analisado com deslocamento igual a zero;

$x[n+k]$ é o sinal a ser analisado com deslocamento igual a k;

M é o valor final para análise.

A autocorrelação mede a relação entre o valor atual de uma variável e seus valores anteriores. Ao calcular a autocorrelação, a saída resultante pode variar de 1 positivo a 1 negativo de acordo com a estatística de correlação tradicional. Uma autocorrelação igual a 1 representa uma correlação positiva perfeita (o sinal e sua versão atrasada são idênticos). Uma autocorrelação igual a -1, por outro lado, representa uma correlação negativa perfeita (o sinal é o inverso negativo de sua versão atrasada/adiantada).

Assim, a autocorrelação mede relações lineares e, mesmo que a autocorrelação seja minúscula, ainda pode haver uma relação não linear entre uma série temporal e uma versão defasada de si mesma. Desta forma, esta ferramenta matemática auxilia na busca para encontrar padrões repetitivos, como a presença de caracteres repetidos na estrutura literária poética.

Tendo em vista a definição da função autocorrelação, sabe-se que o valor de maior autocorrelação equivale ao ponto de maior similaridade do sinal. Desta forma, quando os dois sinais estiverem em fase, os dois sinais encontram-se “sobrepostos”.

Com base nisto, poemas de todas as escolas literárias são analisados a fim de estabelecer uma relação de periodicidade entre os elementos constituintes das estruturas poéticas utilizando a função de autocorrelação.

5.2.1 Análise Gráfica

Em virtude da significante quantidade de gráficos extraídos das estruturas propostas, o intuito principal se torna fazer uma análise comparativa entre as escolas literárias nos extremos da periodicidade esperada entre elas. Ou seja, o Parnasianismo e o Modernismo são as escolas literárias que mais e menos, respectivamente, atendem a necessidade da métrica e ritmos regulares. Desta forma, as composições textuais propostas no pré-processamento serão analisadas conforme as duas escolas.

Assim, dentro do contexto da autocorrelação dos sinais provenientes destes poemas analisados, o intuito é encontrar o maior valor de autocorrelação do sinal para identificar uma maior chance de repetitividade conforme a posição deste elemento no sinal. Os pontos de máxima autocorrelação, então, representam o passo no qual o poema deve apresentar uma repetição, já que, como explicitado, uma maior autocorrelação indica uma maior similaridade.

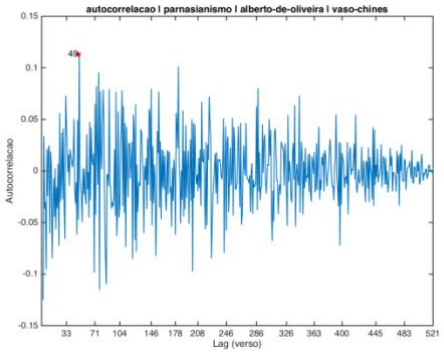
Como proposto, a pré-seleção da estrutura é analisada conforme sete diferentes tipos de pré-processamento pré-estabelecidos. O objetivo é minerar informações provenientes das diferentes estruturas buscando encontrar padrões nos poemas e/ou interpretar o papel de diferentes elementos nos textos para, finalmente, encontrar possíveis relações entre as rimas destes poemas e seus valores de autocorrelação.

Ou seja, pretende-se identificar possíveis repetições em cada verso ou estrofe dos textos selecionados. De tal maneira, a análise contempla de que forma as palavras (e seus valores) contribuem para tal aliteração. Assim, qualquer terminação incomum (como por exemplo mandala/dá-la), pode ser adaptada para uma direta correlação. As Figuras 12 a 25 representam, então, os valores de autocorrelação para cada modificação em função do comprimento total do poema. Na abcissa ainda é possível notar o tamanho de cada verso quando comparado ao tamanho do poema total.

5.2.1.1 Poema original, com nível nulo de modificações, composto por letras maiúsculas e minúsculas, sinais gráficos e acentuação;

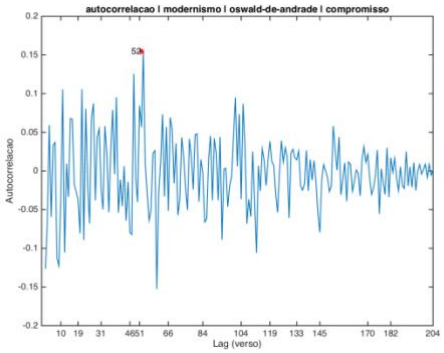
As Figuras 12 e 13, referentes a esta primeira análise, representam os valores de autocorrelação do período completo do texto modificado. Desta forma, nota-se o maior valor de autocorrelação no 49º elemento do texto parnasianista e no 52º elemento do texto modernista. Ou seja, a similaridade dos textos deve ser notada a cada 49 e 52 caracteres dos textos, respectivamente.

Figura 12. Parnasianismo (versão 1)



Fonte: Autoria Própria.

Figura 13. Modernismo (versão 1)



Fonte: Autoria Própria.

Os textos pré-processados para este caso representam a repetitividade de caracteres de acordo com o resultado obtido através da autocorrelação dos elementos dispostos nesta forma de análise. Assim, em amarelo e azul, é possível representar visualmente a periodicidade obtida a cada 49 e 52 elementos dos textos analisados.

Estranho mimo aquele vaso! Vi-o.
Casualmente, uma vez, de um perfumado
Contador sobre o mármor luzidio,
Entre um leque e o começo de um bordado.

Fino artista chinês, enamorado,
Nele pusera o coração doentio
Em rubras flores de um sutil lavrado,
Na tinta ardente, de um calor sombrio.

Mas, talvez por contraste à desventura,
Quem o sabe?... de um velho mandarim
Também lá estava a singular figura;

Que arte em pintá-la! a gente acaso vendo-a,
Sentia um não sei quê com aquele chim
De olhos cortados à feição de amêndoa.

Comprarei
O pincel
Do Douanier
P'ra te pintar
Levo
P'ro nosso lar
O piano periquito
E o Reader's Digest
Pra não temer
Quando morrer
E te deixar
Eu quero nunca te deixar
Quero ficar
Preso ao teu amanhecer

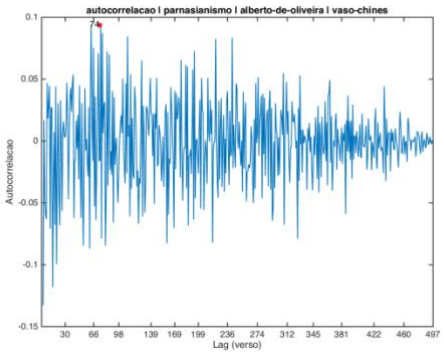
Desta forma, nota-se que os valores de autocorrelação do texto pré-processado não representam necessariamente o padrão esperado para cada rima ou estrofe dos poemas analisados. A rima, neste poema parnasianista, se repete a cada dois versos, diferente do que foi obtido matematicamente.

Para esse caso, versos têm comprimento médio de 36 caracteres. Assim, o esperado era que a autocorrelação resultasse em algum valor aproximadamente igual a 72, representando as rimas obtidas com a versificação do texto, ou 144 para estrofes inteiras.

5.2.1.2 Poema sem letras maiúsculas, substituindo toda e qualquer letra maiúscula por sua equivalente minúscula, assim como caracteres acentuados por seus respectivos sem acentuação, e também eliminando sinais de pontuação;

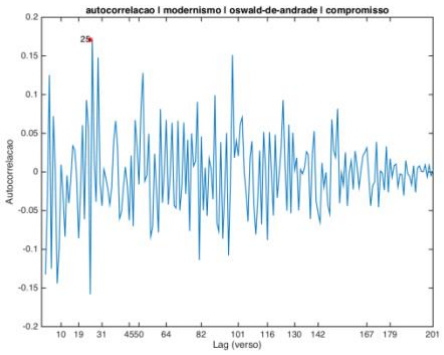
As Figuras 14 e 15, referentes à segunda análise, representam os valores de autocorrelação do período completo do texto modificado. Desta forma, nota-se o maior valor de autocorrelação no 74º elemento do texto parnasianista e no 25º elemento do texto modernista. Ou seja, a similaridade dos textos deve ser notada a cada 74 e 25 caracteres do texto, respectivamente.

Figura 14. Parnasianismo (versão 2)



Fonte: Autoria Própria.

Figura 15. Modernismo (versão 2)



Fonte: Autoria Própria.

Os textos pré-processados para este caso, demonstrados a seguir, representam a repetitividade de caracteres de acordo com o resultado obtido através da autocorrelação dos elementos dispostos nesta forma de análise. Assim, em amarelo e azul, é possível representar visualmente a periodicidade obtida a cada 74 e 25 elementos dos textos analisados.

estranho mimo aquele vaso vio
casualmente uma vez de um perfumado
contador sobre o marmor luzidio
entre um leque e o comeco de um bordado

fino artista chines enamorado
nele pusera o coracao doentio
em rubras flores de um sutil lavrado
na tinta ardente de um calor sombrio

mas talvez por contraste a desventura
quem o sabe de um velho mandarim
tambem la estava a singular figura

que arte em pintala a gente acaso vendoa
sentia um nao sei que com aquele chim
de olhos cortados a feicao de amendoa

comprarei
o pincel
do douanier
pra te pintar
levo
pro nosso lar
o piano periquito
e o readers digest
pra nao temer
quando morrer
e te deixar
eu quero nunca te deixar
quero ficar
preso ao teu amanhecer

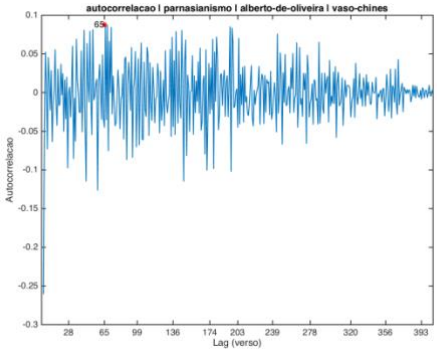
Desta forma, igualmente à versão anterior do poema, nota-se que os valores de autocorrelação do texto pré-processado também não representam necessariamente o padrão esperado para cada rima ou estrofe dos poemas analisados.

Para esse caso, de mesma forma, versos têm comprimento médio de 36 caracteres. Assim, o esperado era que a autocorrelação resultasse em algum valor aproximadamente igual a 72, representando as rimas obtidas com a versificação do texto, ou 144 para estrofes inteiras.

5.2.1.3 Poema sem letras maiúsculas, substituindo toda e qualquer letra maiúscula por sua equivalente minúscula, assim como caracteres acentuados por seus respectivos sem acentuação, e também eliminando sinais de pontuação e caracteres “invisíveis”;

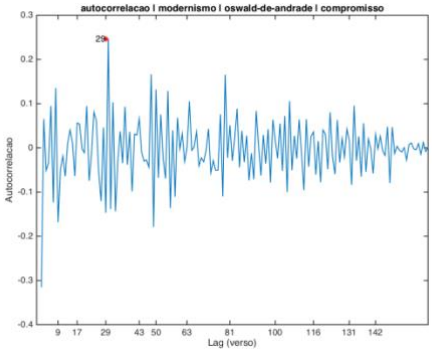
As Figuras 16 e 17, referentes à terceira análise, representam os valores de autocorrelação do período completo do texto modificado. Desta forma, nota-se o maior valor de autocorrelação no 65º elemento do texto parnasianista e no 29º elemento do texto modernista. Ou seja, a similaridade dos textos deve ser notada a cada 65 e 29 caracteres dos textos, respectivamente.

Figura 16. Parnasianismo (versão 3)



Fonte: Autoria Própria.

Figura 17. Modernismo (versão 3)



Fonte: Autoria Própria.

Os textos pré-processados para este caso, demonstrados a seguir, representam a repetitividade de caracteres de acordo com o resultado obtido através da autocorrelação dos elementos dispostos nesta forma de análise. Assim, em amarelo e azul, é possível representar visualmente a periodicidade obtida a cada 65 e 29 elementos dos textos analisados.

estranhomimoaquelevasoviocasualmente
umavezdeumperfumadocontadorsobreo
marmorluzidioentreumlequeeocomecod
eumbordadofinoartistachinesenamorado
nelepuseraocoracaodoentioemrubrasflor
esdeumsutillavradonatintaardentedeumc
alorsombriomastalvezporcontrasteadesv
enturaquemosabedeumvelhomandarimt
ambemlaestavaasingularfiguraqueartee
mpintalaagenteacasovendoasentiaumnao
seiquecomaquelechimdeolhoscortadosaf
eicao deamendoa

comprareiopinceldodouanierpratepint
arlevopronossolaropianoperiquitoeore
adersdigestpranaotemerquandomorre
retedeixareuqueronuncatedeixarquero
ficarpresoaoteuamanhecer

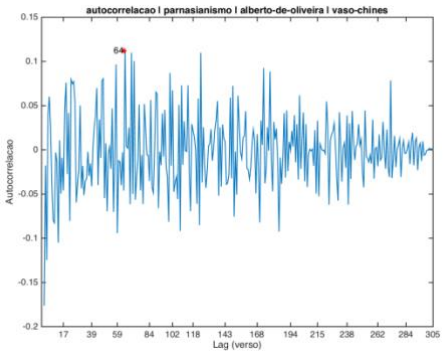
Desta forma, nota-se que os valores de autocorrelação do texto pré-processado também não representam necessariamente o padrão esperado para cada rima ou estrofe dos poemas analisados.

Para esse caso, versos têm comprimento médio de 28 caracteres. Assim, o esperado era que a autocorrelação resultasse em algum valor aproximadamente igual a 56, representando as rimas obtidas com a versificação do texto, ou 112 para estrofes inteiras.

5.2.1.4 Poema sem letras maiúsculas, substituindo toda e qualquer letra maiúscula por sua equivalente minúscula e eliminando vogais e pontuação;

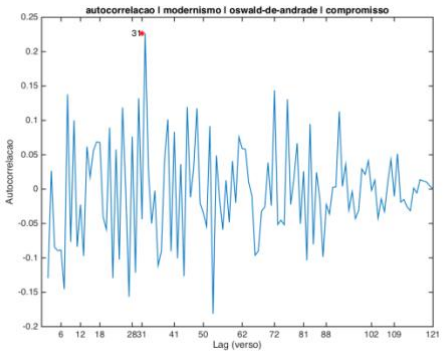
As Figuras 18 e 19, referentes à quarta análise, representam os valores de autocorrelação do período completo do texto modificado. Desta forma, nota-se o maior valor de autocorrelação no 64º elemento do texto parnasianista e no 31º elemento do texto modernista. Ou seja, a similaridade dos textos deve ser notada a cada 64 e 31 caracteres dos textos, respectivamente.

Figura 18. Parnasianismo (versão 4)



Fonte: Autoria Própria.

Figura 19. Modernismo (versão 4)



Fonte: Autoria Própria.

Os textos pré-processados para este caso, demonstrados a seguir, representam a repetitividade de caracteres de acordo com o resultado obtido através da autocorrelação dos elementos dispostos nesta forma de análise. Assim, em amarelo e azul, é possível representar visualmente a periodicidade obtida a cada 64 e 31 elementos dos textos analisados.

strnh mm ql vs v
cslmnt m vz d m prfmd
cntdr sbr mrmr lzd
ntr m lq cmc d m brdd

fn rtst chns nmrd
nl psr crc dnt
m rbrs flrs d m stl lvrd
n tnt rdnt d m clr smbr

ms tlvz pr cntrst dsvntr
qm sb d m vlh mndrm
tmbm l stv snglr fgr

q rt m pntl gnt cs vnd
snt m n s q cm ql chm
d lhs crtds fc d mnd

cmprr
 pncl
d dnr
pr t pntr
lv
pr nss lr
 pn prqt
 rdrs dgst
pr n tmr
qnd mrrr
 t dxr
u qr nnc t dxr
qr fcr
prs t mnhcr

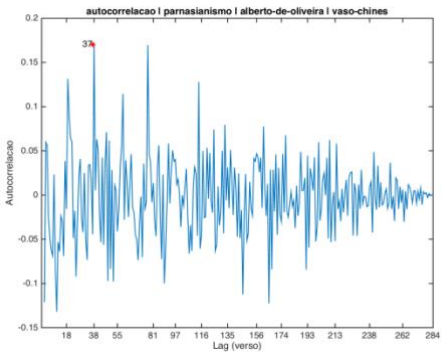
Desta forma, igualmente à versão anterior do poema, nota-se que os valores de autocorrelação do texto pré-processado também não representam necessariamente o padrão esperado para cada rima ou estrofe dos poemas analisados.

Para esse caso, de mesma forma, versos têm comprimento médio de 21 caracteres. Assim, o esperado era que a autocorrelação resultasse em algum valor aproximadamente igual a 42, representando as rimas obtidas com a versificação do texto, ou 84 para estrofes inteiras.

5.2.1.5 Poema sem letras maiúsculas, substituindo toda e qualquer letra maiúscula por sua equivalente minúscula e eliminando consoantes e pontuação;

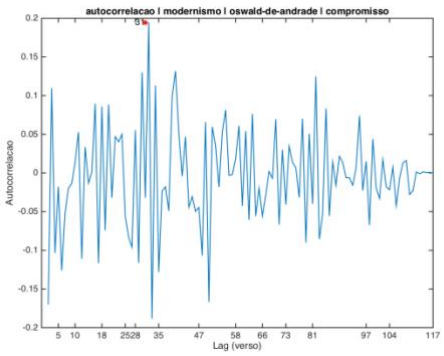
As Figuras 20 e 21, referentes à quinta análise, representam os valores de autocorrelação do período completo do texto modificado. Desta forma, nota-se o maior valor de autocorrelação no 37º elemento do texto parnasianista e no 31º elemento do texto modernista. Ou seja, a similaridade dos textos deve ser notada a cada 37 e 31 caracteres do texto, respectivamente.

Figura 20. Parnasianismo (versão 5)



Fonte: Autoria Própria.

Figura 21. Modernismo (versão 5)



Fonte: Autoria Própria.

Os textos pré-processados para este caso, demonstrados a seguir, representam a repetitividade de caracteres de acordo com o resultado obtido através da autocorrelação dos elementos dispostos nesta forma de análise. Assim, em amarelo e azul, é possível representar visualmente a periodicidade obtida a cada 37 e 31 elementos dos textos analisados.

eao io auee ao io
auaee ua e e u euao
oao oe o ao uiio
ee u eue e o oeo e u oao

io aia ie eaoao
ee uea o oaao oeio
e ua oe e u ui aao
a ia aee e u ao oio

a ae o oae a eeua
ue o ae e u eo aai
ae a eaa a iua iua

ue ae e iaa a ee aao eoa
eia u ao ei ue o auee i
e oo oao a eiao e aeoa

oaei
o ie
o ouaie
a e ia
eo
ro oo a
o iao eiuio
e o eae ie
ra ao ee
uao oe
e e eia
eu ueo ua e eia
ueo ia
eo ao eu aaee

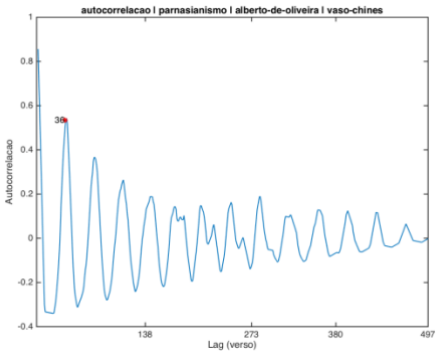
Desta forma, igualmente à versão anterior do poema, nota-se que os valores de autocorrelação do texto pré-processado também não representam necessariamente o padrão esperado para cada rima ou estrofe dos poemas analisados.

Para esse caso, de mesma forma, versos têm comprimento médio de 19 caracteres. Assim, o esperado era que a autocorrelação resultasse em algum valor aproximadamente igual a 38, representando as rimas obtidas com a versificação do texto, ou 76 para estrofes inteiras.

5.2.1.6 Poema sem letras maiúsculas, substituindo toda e qualquer letra maiúscula por sua equivalente minúscula e convertendo 75% de cada verso para valor nulo;

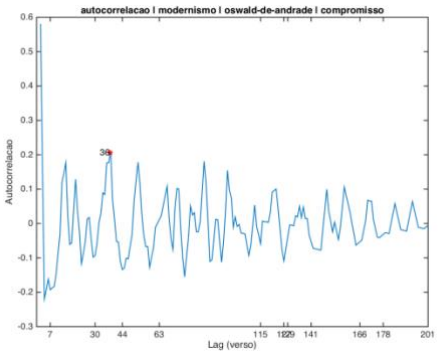
As Figuras 22 e 23, referentes à sexta análise, representam os valores de autocorrelação do período completo do texto modificado. Desta forma, nota-se o maior valor de autocorrelação no 36º elemento do texto parnasianista e no 36º elemento do texto modernista. Ou seja, a similaridade dos textos deve ser notada a cada 36 caracteres dos textos, respectivamente.

Figura 22. Parnasianismo (versão 6)



Fonte: Autoria Própria.

Figura 23. Modernismo (versão 6)



Fonte: Autoria Própria.

Os textos pré-processados para este caso, demonstrados a seguir, representam a repetitividade de caracteres de acordo com o resultado obtido através da autocorrelação dos elementos dispostos nesta forma de análise. Assim, em amarelo e azul, é possível representar visualmente a periodicidade obtida a cada 36 elementos dos textos analisados.

vaso vio
perfumado
r luzidio
m bordado

enamorado
ao doentio
l lavrado
r sombrio

esventura
mandarim
ar figura

so vendoa
uele chim
e amendoa

arei
ncel
nier
ntar
levo
lar
uito
gest
emer
rrer
ixar
ixar
icar
ecer

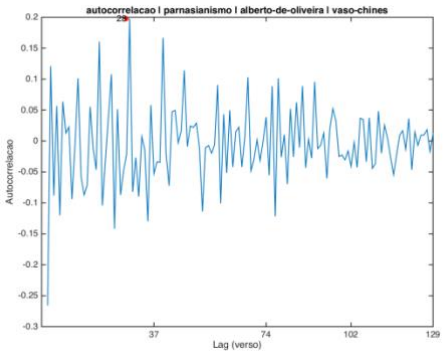
Desta forma, nota-se que os valores de autocorrelação do texto pré-processado também não representam necessariamente o padrão esperado para cada rima ou estrofe dos poemas analisados.

Para esse caso, de mesma forma, versos têm comprimento médio de 36 caracteres. Assim, o esperado era que a autocorrelação resultasse em algum valor aproximadamente igual a 72, representando as rimas obtidas com a versificação do texto, ou 144 para estrofes inteiras.

5.2.1.7 Poema sem letras maiúsculas, substituindo toda e qualquer letra maiúscula por sua equivalente minúscula e eliminando 75% de cada verso;

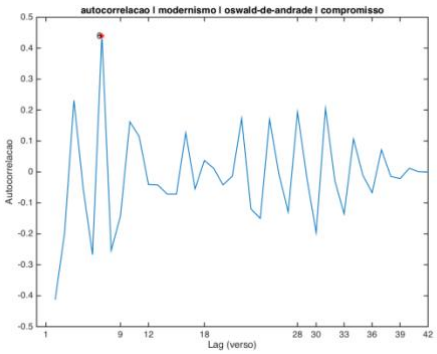
As Figuras 24 e 25, referentes à segunda análise, representam os valores de autocorrelação do período completo do texto modificado. Desta forma, nota-se o maior valor de autocorrelação no 28º elemento do texto parnasianista e no 6º elemento do texto modernista. Ou seja, a similaridade dos textos deve ser notada a cada 28 e 6 caracteres do texto, respectivamente.

Figura 24. Parnasianismo (versão 7)



Fonte: Autoria Própria.

Figura 25. Modernismo (versão 7)



Fonte: Autoria Própria.

Os textos pré-processados para este caso, demonstrados a seguir, representam a repetitividade de caracteres de acordo com o resultado obtido através da autocorrelação dos elementos dispostos nesta forma de análise. Assim, em amarelo e azul, é possível representar visualmente a periodicidade obtida a cada 28 e 6 elementos dos textos analisados.

vaso vioperfumador luzidio m bordado
enamoradoao doentiol lavrador sombrio
esventuramandarimar figura
so vendoauele chime amendoa

areincelnierntarlevolaruito
gestemerrrerixarixaricarecer

Desta forma, nota-se que os valores de autocorrelação do texto pré-processado também não representam necessariamente o padrão esperado para cada rima ou estrofe dos poemas analisados.

Para esse caso, de mesma forma, versos têm comprimento médio de 9 caracteres. Assim, o esperado era que a autocorrelação resultasse em algum valor aproximadamente igual a 18, representando as rimas obtidas com a versificação do texto, ou 56 para estrofes inteiras.

5.3 RESULTADOS

Como resultado da análise gráfica proveniente da autocorrelação dos poemas estudados, não é possível relacionar tais valores de autocorrelação obtidos com as padronizações dos textos. Ainda que tenha sido demonstrado apenas um único poema parnasianista (considerando sua notável identidade métrica), a invalidez do estudo foi igualmente comprovada para os outros poemas analisados.

Cabe ressaltar como melhor resultado o caso onde há a substituição dos primeiros 75% caracteres de cada verso. Entretanto, este resultado é errôneo, pois a substituição de caracteres pelo elemento nulo (zero) faz com que a autocorrelação seja evidentemente maior, já que 75% do texto será composto por este valor. Ainda assim, quando há a eliminação de 75% de cada verso, obtemos um valor maior quando comparado aos outros métodos de pré-processamento.

É possível demonstrar os baixos valores obtidos com o estudo através da Tabela 4, representando os valores dos picos de autocorrelação por escola literária.

Tabela 4. Valores de pico de autocorrelação por escola literária.

Escola Literária	Período Fundamental (médio)	Autocorrelação (média)						
		Análise 1	Análise 2	Análise 3	Análise 4	Análise 5	Análise 6	Análise 7
Arcadismo	120	0,1020	0,0964	0,1152	0,1192	0,1360	0,8214	0,2652
Barroco	52	0,1181	0,1131	0,1426	0,1314	0,1501	0,8432	0,2298
Modernismo	120	0,1254	0,1376	0,1707	0,1701	0,1693	0,7655	0,2514
Parnasianismo	96	0,1115	0,1103	0,1248	0,1372	0,1407	0,8462	0,2421
Realismo	129	0,0969	0,0985	0,1276	0,1222	0,1300	0,8372	0,2221
Romantismo	72	0,1030	0,0928	0,1085	0,1128	0,1224	0,8092	0,2228
Simbolismo	67	0,1163	0,1123	0,1245	0,1284	0,1438	0,8457	0,2414

6 CONCLUSÃO

A motivação para este trabalho foi investigar a viabilidade do reconhecimento de padrões em poemas. Não foi pretendido tirar conclusões concretas para esse trabalho, já que se trata de uma mineração de dados. É importante frisar que diversas experiências adicionais ainda são necessárias para a obtenção de resultados concretos. Mais importante, o principal objetivo deste trabalho foi iniciar a pesquisa no campo do reconhecimento de padrões em textos poéticos através da autocorrelação, fornecendo pensamentos preliminares, experimentos e resultados.

Assim, o principal objetivo do trabalho leva em conta a representação de padrões em sua totalidade, quando apenas um conjunto de amostras está disponível. Entretanto, a implantação da análise não se torna plausível neste caso. Os poetas tomam decisões sobre o impacto relativo de cada palavra no ritmo total da linha e estruturam o poema de acordo com os efeitos pretendidos que, às vezes, são fortuitos. Desta forma, problemas podem ser enfrentados durante a execução de uma análise referente ao medidor poético.

Primeiro, como sugerido acima, a interpretação de um poema por um leitor afetará a produção de um padrão métrico para uma linha específica. Outra área de dificuldade em sistemas de análise métrica é a notação. O estresse textual é mais frequentemente listado como um valor binário (forte ou fraco) e, mesmo quando esses valores são entendidos como valores relacionais e não absolutos, pode haver dificuldades nas suas quantificações.

Tais questões são legítimas, pois a poesia é um ato de fala altamente ensaiado. Entretanto cabe, aqui, a sugestão de análise para fonemas, ao invés de caracteres individualmente considerados. Os fonemas, como são as menores unidades sonoras do sistema fonológico, podem representar uma notável padronização ao analisarmos este tipo de texto.

Por fim, o presente trabalho estabelece o trampolim para o reconhecimento de poemas e as aplicações da autocorrelação nesta análise, ainda que as técnicas utilizadas neste trabalho não permitiram a identificação proposta, devido seus baixos valores obtidos através da autocorrelação. Possíveis desafios são destacados no estudo com o intuito de iluminar o aspecto técnico do reconhecimento de textos e propor aos trabalhos futuros novas formas de extração de conhecimento provenientes de estruturas literárias. Estratégias mais adaptativas, como a aplicação de inteligência artificial ou machine learning, podem ser uma alternativa interessante a ser explorada.

REFERÊNCIAS

ACKOFF, Russell Lincoln. From data to wisdom. **Journal of applied systems analysis,** Stanford, v. 16, p. 3-9, mar. 1989.

AGAZZI, Evandro; MINAZZI, Fabio. **Science and ethics**: The axiological contexts of science. Bruxelas: P.I.E Peter Lang, 2008. 299 p.

APOSTILA DE TEORIA – GRADUAÇÃO, ENGENHARIA DE TELECOMUNICAÇÕES. **Apostila de teoria para processamento digital de sinais**. Disponível em: <http://www.telecom.uff.br/~delavega/public/dsp/apostila_teo_dsp.pdf>. Acesso em: 30 abr. 2017.

BEHRENS, John. **Principles and Procedures of Exploratory Data Analysis.** American Psychological Association, Arizona, v. 2, n. 2, p. 131-160, jun. 2017. Disponível em: <http://cll.stanford.edu/~willb/course/behrens97pm.pdf>. Acesso em: 21 mai. 2017.

BARBOSA, Frederico; SANTOS, Elaine Cuenca. **Modernismo na literatura brasileira.** 1 ed. Curitiba: IESDE Brasil, 2009. 280 p.

BASTOS, Alcmeno. **Poesia brasileira e estilos de época.** 2 ed. Rio de Janeiro: 7Letras, 2004. 326 p.

BERNARDI, Francisco. **As bases da literatura brasileira:** Histórias, autores, textos e testes. 1 ed. Porto Alegre: AGE Editora, 1999. 217 p.

BOULET, Benoit. **Fundamentals of signals and systems.** Massachusetts: Charles River Media, 2006. 670 p.

CASTELLO, José Aderaldo. **A literatura brasileira:** Origens e Unidade (1500 - 1960). 1 ed. São Paulo: Editora da Universidade de São Paulo, 2004. 432 p.

CHEVEIGNÉ, Alain de; KAWAHARA, Hideki. **YIN, a fundamental frequency estimator for speech and music**. Acoustical Society of America, Boston, v. 111, n. 4, p. 1917-1930, mai. 2002. Disponível em: <http://recherche.ircam.fr/equipes/pcm/cheveign/ps/2002_JASA_YIN_proof.pdf>. Acesso em: 03 mai. 2017.

CIOS, K. J. et al. **Data mining: a knowledge discovery approach**. Denver: Springer, 2007. 605 p.

DEWEY, Tonya Kim; FROG. **Versatility in versification**: Multidisciplinary approaches to metrics. New York: Peter Lang, 2009. 301 p.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **American Association for Artificial Intelligence**, Michigan, v. 17, n. 3, p. 37-54, mar. 1996. Disponível em: <https://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>. Acesso em: 21 mai. 2017.
GARTNER. **Data mining**. Disponível em: <http://www.gartner.com/it-glossary/data-mining>. Acesso em: 06 set. 2016.

GAYDECKI, Patrick. **Foundations of digital signal processing: Theory, algorithms and hardware design.** 1 ed. Londres: The Institution of Electrical Engineers, 2004. 461 p.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining: um guia prático**. 4 ed. Rio de Janeiro: Elsevier, 2005.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques.** 3 ed. Illinois: Elsevier, 2009.

HAND, David; MANNILA, Heikki; SMYTH, Padhraic. **Principles of data mining**. 1 ed. Massachusetts: MIT, 2001. 449 p.

LEVY, Pierre. **A inteligência coletiva.** 5 ed. São Paulo: Loyola, 2007.

LIMA, Ivan De. **Literature, leitura e compreensão de textos**. São Paulo: Clube de autores, 2009. 87 p.

MAIMON, Oded; ROKACH, Lior. **Data mining and knowledge discovery handbook**. 2 ed. Tel-Aviv: Springer, 2005. 1284 p.

MARTINS, Gilberto de Andrade. **Estudo de Caso: uma estratégia de pesquisa**. 2 ed. São Paulo. Atlas, 2008.

MATIA, Kátia Caroline de. **Poesia expandida: A escrita poética no ciberespaço**. 2013. 111f. **Dissertação de Mestrado - Universidade Estadual de Maringá, Letras/Estudos Literários, Maringá, 2013**. Disponível em: <http://www.ple.uem.br/defesas/pdf/kcmatia.pdf>. Acesso em: 06 set. 2016.

MICHAELIS. **Moderno Dicionário da Língua Brasileira**. Disponível em: <http://michaelis.uol.com.br/busca?r=0&f=0&t=0&palavra=literatura>. Acesso em: 28 mar. 2017.

MICHAELIS. **Moderno Dicionário da Língua Brasileira**. Disponível em: <http://michaelis.uol.com.br/busca?r=0&f=0&t=0&palavra=verso>. Acesso em: Acesso em: 28 mar. 2017.

MICHAELIS. **Moderno Dicionário da Língua Brasileira**. Disponível em: <http://michaelis.uol.com.br/busca?r=0&f=0&t=0&palavra=métrica>. Acesso em: Acesso em: 28 mar. 2017.

MICHAELIS. **Moderno Dicionário da Língua Brasileira**. Disponível em: <http://michaelis.uol.com.br/busca?r=0&f=0&t=0&palavra=poesia>. Acesso em: Acesso em: 28 mar. 2017.

MOISÉS, Massaud. **A literatura brasileira através dos textos**. 25 ed. São Paulo: Cultrix, 1999. 686 p.

MOISÉS, Massaud. **Dicionário de termos literários**. 1 ed. São Paulo: Cultrix, 1974. 523 p.

MORITA, Kiyoshi. **Applied fourier transform**. Japão: Ohmsha, 1995. 439 p.

OLIVEIRA, Silvana. **Realismo na literatura brasileira**. 1 ed. Curitiba: IESDE Brasil, 2008. 204 p.

OLSON, David. **Descriptive data mining**. 1 ed. Lincoln: Springer, 2017. 116 p.

OLSON, David L.; WU, Desheng. **Predictive data mining**. 1 ed. Lincoln: Springer, 2017. 99 p.

OLSON, Harry F.. **Music, physics and engineering**. 2 ed. New Jersey: Dover Publications, 1967. 468 p.

OORD, Aaron Van Den; DIELEMAN, Sander; SCHRAUWEN, Benjamin. Deep content-based music recommendation. **Advances in Neural Information Processing Systems**, Lake Tahoe, n. 26, p. 2543-2651, jul. 2013. Disponível em: <https://papers.nips.cc/paper/5004-deep-content-based-music-recommendation.pdf>. Acesso em: 21 set. 2016.

PRANDONI, Paolo; VETTERLI, Martin. **Signal processing for communications**. 1 ed. Lausanne: CRC Press, 2008. 371 p.

PRIEMER, Roland. **Introductory signal processing.** Illinois: World Scientific, 1991. 739 p.

NAHIN, Paul J.. **Dr. Euler's fabulous formula: cures many mathematical ills**. 8 ed. New Jersey: Princeton University, 2006. 381 p.

RABINER, Lawrence R.; GOLD, Bernard. **Theory and application of digital signal processing**. New Jersey: Prentice-Hall, 1978. 510 p.

RIBEIRO, A. E. et al. **Leitura e escrita em movimento**. São Paulo: Peirópolis, 2012. 8 p.

ROWLEY, Jennifer. **The wisdom hierarchy: representations of the dikw hierarchy**. Journal of information science, Chicago, v. 33, n. 2, p. 163-180, fev. 2007. Disponível em: <http://wisdomresearch.org/forums/storage/26/220/rowley_jis_042007.pdf>. Acesso em: 21 mai. 2017.

SILVA, Antonio Manoel Dos Santos; SANT'ANNA, Romildo. **Literaturas de língua portuguesa:** Brasil. São Paulo: Arte e Ciência, 2007. 288 p.

SMITH, Steven W.. **The scientist and engineer's guide to digital signal processing**. 2 ed. San Diego: California Technical Publishing, 1999. 650 p.

STRANNEBY, Dag. **Digital signal processing: DSP and applications.** Massachusetts: Newnes, 2001. 229 p.

THE COMPLETE TABLE OF ASCII CHARACTERS, LETTERS, CODES, SYMBOLS AND SIGNS. **ASCII table, ASCII codes**. Disponível em: <http://www.theasciicode.com.ar>. Acesso em: 25 mai. 2017.

THE UNICODE CONSORTIUM. **The unicode standard**. Disponível em: <http://www.unicode.org/versions/unicode6.0.0/unicodestandard-6.0.pdf>. Acesso em: 26 mai. 2017.

TIZHOOSH, Hamid R.; SAHBA, Farhang; DARA, Rozita. **Poetic Features for Poem Recognition: A Comparative Study**. Journal of Pattern Recognition Research, v. 3, n. 1, p. 24-39, set. 2008.

TRINGALI, Dante. **Escolas literárias:** São Paulo: Musa, 1994. 246 p.

TURCO, Lewis**. The New Book of Forms: A Handbook of Poetics**. Hanover and London: University Press, 1986.

VASEGHI, Saeed V.. **Advanced digital signal processing and noise reduction**. 4 ed. Singapura: Wiley, 2008. 513 p.

WILLINGER, W. et al. **Scaling phenomena in the Internet: Critically examining criticality. Proceedings of the National Academy of Sciences**, Irvine, v. 99, n. 1, p. 2573-2580, fev. 2006. Disponível em: <http://www.pnas.org/content/99/suppl_1/2573.full.pdf>. Acesso em: 05 set. 2016.

APÊNDICES

APÊNDICE I – CÓDIGO DESENVOLVIDO

```
mov_input_sel = input(' 1: arcadismo \n 2: barroco \n 3: modernismo \n 4: parnasianismo \n 5:
realismo \n 6: romantismo \n 7: simbolismo \n <enter> \n >> ', 's');
mov_input = mov_select(mov_input_sel);

%author_input = input('author name or <enter> \n >> ', 's');
author_input = '';

method_gen_sel = input(' 0: personalize \n 1: original poem \n 2: lowercase, wo_acc, wo_pont \n 3:
lowercase, wo_acc, wo_pont, wo_invis \n 4: lowercase, wo_acc, wo_pont, wo_vowels \n 5: lowercase,
wo_acc, wo_pont, wo_conso \n 6: lowercase, wo_acc, wo_pont, wo_char \n 7: lowercase, wo_acc,
wo_pont, wo_char_zeros \n>> ', 's');
[method_s, method_a, method_p, method_i, method_v, method_c, method_l, method_l2] =
method_select(method_gen_sel);

poem_conv_all = [];
length_poem_all = [];
contagem_todos = [];
contagem_poema = 0;

level1 = dir();
for idx1 = 1:length(level1)
    if strcmp(level1(idx1).name, mov_input)
        % entrando no nivel de escola literaria
        disp(sprintf('movement [%s]', level1(idx1).name));

        level2 = dir(mov_input);
        for idx2=1:length(level2)
            % seleciona todos os poemas de todos os autores
            if isempty(author_input) || strcmp(level2(idx2).name, author_input)
                if level2(idx2).isdir == false
                    continue
                end
                if strcmp(level2(idx2).name, '.') || strcmp(level2(idx2).name, '..') ||
strcmp(level1(idx1).name, '.DS_Store')
                    continue
                end;

                % entrando no nivel dos autores dessa escola
                disp(sprintf(' author [%s]', level2(idx2).name));

                fullname = fullfile(level1(idx1).name, level2(idx2).name);
                level3 = dir(fullname);
                for idx3=1:length(level3)
                    if level3(idx3).isdir == true
                        continue
                    end
                    if strcmp(level2(idx2).name, '.') || strcmp(level2(idx2).name, '..') ||
strcmp(level3(idx3).name, '.DS_Store')
                        continue
                    end;

                    % entrando no nivel dos poemas desse autor
                    disp(sprintf('  poem [%s]', level3(idx3).name));
                    fullname = fullfile(...
                        level1(idx1).name, level2(idx2).name, level3(idx3).name);
                    currentCharacterEncoding = slCharacterEncoding('ISO-8859-1');
                    fid = fopen(fullname, 'r');
                    poem = fscanf(fid, '%c');
                    fclose(fid);

                    % input do usuario para configurar metodo de analise
                    poem_conv = method_input(poem, method_s, method_a, method_p, method_i, method_v,
method_c, method_l, method_l2);

                    contagem_poema = contagem_poema + 1;
                    length_poema(contagem_poema) = length(poem_conv);
                    length_poem = length(poem_conv);
                    length_poem_all = [length_poem_all, length_poem];
                    length_poem_mean = mean(length_poem_all);

                    % concatena todos poemas em um unico poema
                    poem_conv_all = [poem_conv_all, poem_conv];
```

```
                    % recodifica simbolos_todos baseado na frequencia de char
                    [symbol_code_all, frequency_all, characters_all] =
get_symbol_frequency(poem_conv_all);

                    % histograma da escola
                    h = figure;
                    set(gcf, 'Color', 'white');
                    hist = bar(frequency_all/sum(frequency_all));
                    xlabel('Simbolos');
                    ylabel('Frequencia relativa');
                    set(get(hist , 'Parent'), 'XTickLabel', symbol_code_all);
                    set(get(hist , 'Parent'), 'XTick', 1:length(frequency_all));
                    set(get(hist , 'Parent'), 'XTickLabelRotation', 45);
                    filename = sprintf('Histograma - %s', level1(idx1).name);
                    title(filename);
                    saveas(gcf, filename, 'png');

                end
            end
        end

    end
end

% recodifica simbolos_todos baseado na frequencia de char
[symbol_code_all, frequency_all, characters_all] = get_symbol_frequency(poem_conv_all);

% gera novo codigo de acordo com a frequencia dos simbolos
index_norm = 1;
new_symbol_code_all = (frequency_all/max(frequency_all)).^index_norm;

% repete recodificacao de todos os poemas para fazer analises por poema
level1 = dir();
for idx1 = 1:length(level1)
    if strcmp(level1(idx1).name, mov_input)
        % entrando no nivel de escola literaria
        disp(sprintf('movement [%s]', level1(idx1).name));

        level2 = dir(mov_input);
        for idx2=1:length(level2)
            % seleciona todos os poemas de todos os autores
            if isempty(author_input) || strcmp(level2(idx2).name, author_input)
                if level2(idx2).isdir == false
                    continue
                end
                if strcmp(level2(idx2).name, '.') || strcmp(level2(idx2).name, '..') ||
strcmp(level1(idx1).name, '.DS_Store')
                    continue
                end;

                % entrando no nivel dos autores dessa escola
                disp(sprintf(' author [%s]', level2(idx2).name));

                fullname = fullfile(level1(idx1).name, level2(idx2).name);
                level3 = dir(fullname);
                for idx3=1:length(level3)
                    if level3(idx3).isdir == true
                        continue
                    end
                    if strcmp(level2(idx2).name, '.') || strcmp(level2(idx2).name, '..') ||
strcmp(level3(idx3).name, '.DS_Store')
                        continue
                    end;

                    % entrando no nivel dos poemas desse autor
                    disp(sprintf('  poem [%s]', level3(idx3).name));
                    fullname = fullfile(...
                        level1(idx1).name, level2(idx2).name, level3(idx3).name);
                    currentCharacterEncoding = slCharacterEncoding('ISO-8859-1');
                    fid = fopen(fullname, 'r');
                    poem = fscanf(fid, '%c');
                    fclose(fid);

                    [pathstr,name,ext] = fileparts(level3(idx3).name);

                    % input do usuario para configurar metodo de analise
                    poem_conv = method_input(poem, method_s, method_a, method_p, method_i, method_v,
method_c, method_l, method_l2);

                    % poem_conv para poema individual
                    poem_recod = recode(poem_conv, symbol_code_all, new_symbol_code_all);
```

```
                    % analise sem considerar recodificacao
                    %poem_recod = poem_conv;

                    % remove nivel dc
                    poem_recod = poem_recod - mean(poem_recod);

                    % retorna indices do break line
                    char_break_idx = find(poem_conv == 10);
                    char_break_idx = [char_break_idx, length(poem_conv)];
                    char_break = poem_conv(char_break_idx);
                    char_break = char_break(1);

                    % retorna indices do espaco
                    char_space_idx = find(poem_conv == 32);
                    char_space = poem_conv(char_space_idx);

                    [num_strophes, lim_strophes, num_verses, lim_verses,length_verses] =
count_attributes(poem_conv, char_break);

                    % quando nao ha invisiveis
                    poem_conv_dob = double(poem);
                    invisibles = [10, 32];
                    invisibles_char_mb = ismember(poem_conv, invisibles);
                    invisibles_char_mb_dob = ismember(poem_conv_dob, invisibles);
                    [num_strophes_dob, lim_strophes_dob, num_verses_dob, lim_verses_dob] =
count_attributes(poem_conv_dob, 10);

                    % conta invisiveis para reduzir de lim_verses
                    count_invisibles = cumsum(invisibles_char_mb_dob);
                    verses_invisibles = lim_verses_dob(1:end);
                    result_verses_invisibles = [count_invisibles(verses_invisibles(1)),
diff(count_invisibles(verses_invisibles))];
                    strophes_invisibles = lim_strophes_dob(1:end);
                    result_strophes_invisibles = [count_invisibles(strophes_invisibles(1)),
diff(count_invisibles(strophes_invisibles))];
                    lim_verses_invisibles = lim_verses_dob - result_verses_invisibles;
                    lim_strophes_invisibles = lim_strophes_dob - result_strophes_invisibles;

                    if strcmp(method_i, 'yes')
                        lim_verses = lim_verses_invisibles;
                        lim_strophes = lim_strophes_invisibles;
                    else
                        lim_verses;
                    end

                    % correlacao
                    length_poem = length(poem_recod);
                    correlations = xcorr(poem_recod, 'coeff');

                    % descobre max e idx de valores
                    xcorr_values = correlations(length_poem+1:end);
                    [sortedValues,sortIndex] = sort(xcorr_values(:),'descend');
                    maxIndex = sortIndex(1);
                    max_xcorr = xcorr_values(maxIndex);

                    % interpolation
                    xcorr_values = interp1(2:length(poem_recod), xcorr_values,
1:1:length(poem_recod));

                    % plot do pico das autocorrelacoes
                    corr_plot = plot(xcorr_values);
                    hold on;
                    corr_plot = plot(maxIndex, max_xcorr, 'r*');
                    strmax = [' ',num2str(maxIndex)];
                    text(maxIndex, max_xcorr, strmax,'HorizontalAlignment','right');

                    xlabel('Lag (verso)');
                    xlim([0 length(poem_conv)]);
                    set(gca,'XTick',lim_verses);
                    ylabel('Autocorrelacao');
                    hold off;
                    filename = sprintf('autocorrelacao | %s | %s | %s', level1(idx1).name,
level2(idx2).name, name);
                    title(filename);
                    filename_1 = sprintf('autocorrelacao | %s | %s | %s | %s | %i', method_gen_sel,
level1(idx1).name, level2(idx2).name, name, index_norm);
                    saveas(gcf, filename_1, 'png');

                end
            end
        end
```

```
    end
end
```

APÊNDICE II – FREQUÊNCIA DE CARACTERES POR ESCOLA

Caractere	Descrição	Arcadismo	Barroco	Modernismo	Parnasianismo	Realismo	Romantismo	Simbolismo
10	[Quebra de Linha]	1407	565	358	690	662	1533	392
32	[Espaço]	6634	4987	1457	3094	3674	5968	2635
33	!	39	8	15	94	116	269	60
34	"	6	22	0	16	17	8	8
39	'	36	28	16	6	2	58	6
40	(	4	9	1	1	3	1	0
41	)	4	8	1	1	3	1	0
42	*	3	0	0	0	0	0	0
44	,	974	700	88	538	515	691	385
45	-	80	45	41	81	86	201	60
46	.	351	186	121	258	431	723	354
58	:	85	35	6	15	17	34	15
59	;	127	35	1	42	12	80	0
63	?	53	31	1	16	16	81	12
65	A	209	140	69	67	70	182	59
66	B	13	23	4	20	12	17	9
67	C	104	120	18	44	40	91	61
68	D	146	110	44	60	68	164	77
69	E	184	141	59	117	115	220	84
70	F	39	31	8	14	19	33	29
71	G	34	11	7	6	11	25	6
72	H	20	17	1	7	8	10	17
73	I	27	15	1	7	7	27	14
74	J	29	16	3	1	7	14	3
76	L	55	38	14	11	16	26	15
77	M	139	99	25	22	118	61	29
78	N	121	65	50	37	64	127	44
79	O	121	60	35	29	43	107	24
80	P	135	111	34	35	74	71	35
81	Q	114	94	28	42	62	122	35
82	R	33	46	9	17	11	21	6
83	S	118	72	20	41	61	119	35
84	T	93	53	35	35	53	65	19
85	U	32	17	0	6	26	20	3
86	V	61	37	16	37	37	28	29
88	X	0	1	0	0	0	1	0
90	Z	2	4	0	0	4	0	0
91	[	0	2	0	0	0	0	0
93	]	0	2	0	0	0	0	0
97	a	4009	2786	1042	2064	2396	3839	1947
98	b	366	247	102	179	206	373	149
99	c	758	657	230	358	499	693	466
100	d	1349	1104	404	746	832	1414	671
101	e	3940	3001	939	1871	2285	3564	1636
102	f	351	320	69	171	186	281	163
103	g	431	284	95	177	214	374	165
104	h	389	207	100	141	272	333	176
105	i	1744	1326	474	812	1084	1631	793
106	j	116	69	15	40	57	111	32
107	k	0	0	7	0	0	0	0
108	l	854	633	186	530	543	931	515
109	m	1411	1009	386	754	925	1475	599
110	n	1566	1204	441	796	918	1597	798
111	o	3973	2717	837	1689	2002	3450	1460
112	p	636	510	135	284	387	567	236
113	q	325	301	79	149	214	260	92
114	r	2428	1803	558	1055	1176	2215	1051
115	s	3215	1837	585	1236	1407	2528	1493
116	t	1525	1209	339	693	704	1273	599
117	u	1583	1062	356	688	929	1483	568
118	v	540	407	119	265	275	495	202
120	x	51	33	12	31	41	40	20
121	y	0	0	0	0	1	0	0
122	z	131	103	34	100	100	150	64
192	À	7	8	1	3	1	6	2
193	Á	7	1	0	0	1	3	1
194	Â	0	0	0	0	0	1	0
195	Ã	0	0	0	0	0	7	0
201	É	6	6	1	5	11	10	9
202	Ê	0	0	0	0	1	0	0
205	Í	0	0	0	1	0	0	0
210	Ò	0	0	0	0	0	1	0
211	Ó	3	8	0	5	0	7	5
212	Ô	0	0	0	0	4	0	0
218	Ú	0	0	0	1	0	0	0
224	à	23	12	4	26	13	25	12
225	á	111	142	45	57	99	153	43
226	â	12	12	2	5	5	14	18
227	ã	275	210	76	76	203	235	69
231	ç	137	87	41	51	101	151	38
233	é	139	60	27	54	88	121	65
234	ê	57	34	18	34	44	54	24
237	î	135	39	17	43	31	40	59
239	í	0	2	0	0	0	0	0
243	ó	88	76	19	29	31	46	44
244	ô	10	20	3	3	6	10	5
245	õ	24	14	4	15	16	15	24
250	ú	30	22	10	15	12	26	16