**ANÁLISE DE AGRUPAMENTO APLICADO A ESTUDOS DE CORPUS EM MÚSICA POPULAR BRASILEIRA**

**Eixo temático geral**: Teoria e análise da música popular

Este trabalho detalha um recorte específico de um projeto de grande abrangência – intitulado Projeto MPB (AUTOR 1; AUTOR 2, 2024) – coordenado pelos presentes autores e iniciado ao final de 2022. O projeto tem como objetivo geral o mapeamento do estilo compartilhado pelos compositores de uma – pelos autores denominada – Prática Comum da MPB, através de estudos de *corpus*. Consequentemente, busca-se também a evidenciação sistemática das particularidades estilísticas individuais dos compositores selecionados para alimentar os *corpora* de análise. Na fase inicial do projeto, 10 nomes foram considerados: Tom Jobim, Ivan Lins, Chico Buarque, Edu Lobo, Caetano Veloso, Djavan, João Bosco, Milton Nascimento, Gilberto Gil e Rita Lee, e cada um dos seus respectivos *corpora* é formado por 50 músicas, analisadas considerando modelos teóricos e metodologias (majoritariamente computacionais) especialmente desenvolvidos para as tarefas. A análise abrange as estruturas melódicas (alturas e ritmo) e harmônicas das músicas, bem como a relação entre melodia e harmonia. Atualmente, após a finalização do processo de consolidação dos 10 *corpora,* o projeto se encontra em fase de análise e avaliação da massa de dados produzidos, o que compreende tanto aspectos qualitativos (avaliação teórico-musical) quanto quantitativos (avaliação estatística). O presente trabalho busca detalhar uma das iniciativas desse último aspecto, que propõe, em sua essência, o agrupamento (ou *clustering*, no original em inglês do termo) dos *corpora* de acordo com afinidades estilísticas, a partir de uma abordagem analítica dos dados através de ferramentas estatísticas específicas. Mais especificamente, a *análise de agrupamento* (IZBICKI; SANTOS, 2020, p. 219) representa um conjunto de técnicas estatísticas e computacionais empregadas para encontrar padrões de similaridades em massas de dados observados. No presente trabalho, deseja-se encontrar tais padrões dentro do grupo dos *corpora* analisados na primeira fase, nos quais informações rítmicas, melódicas, harmônicas e de relação entre notas melódicas e os acorde que as harmonizam são codificadas respectivamente através de histogramas de frequências das seguintes variáveis (AUTOR 2, 2022; 2023): *r*-letras (padrões rítmicos minimais, codificados a partir de um alfabeto de 26 possibilidades), *c*-letras (padrões de contorno melódico minimais, codificados a partir de um alfabeto de sete possibilidades), perfil métrico (corresponde à distribuição de ataques rítmicos dentro de uma grade micrométrica que subdivide cada pulso em 12 partes iguais), tipos acordais (abstrações de acordes, desconsiderando fundamentais e componentes concretos e que preservam apenas suas qualidades), categorias funcionais (as diversas atribuições funcionais que acordes podem assumir em um determinado contexto tonal, p. ex., V, V/VI, bIII7M etc.), classes funcionais (conjuntos homogêneos que reúnem categorias funcionais por afinidades, p. ex., Diatônicos, Empréstimos etc.), e de notas-funções (descrevem relações entre melodia e harmonia em âmbitos locais e globais). A partir desses objetos e considerando cada domínio específico de análise, calculam-se matrizes de distâncias dois a dois entre todos os possíveis pares de *corpora*, através da *divergência de Kullback-Leibler* (COVER; THOMAS, 2006, p. 19), uma ferramenta da Teoria da Informação que nos permite quantificar a dissimilaridade entre dois histogramas *p* e *q* através da fórmula

onde as quantidades e representam as probabilidades de alocação dos histogramas e (referente a cada *corpus* dentro de determinado domínio de análise) na categoria (qualidades possíveis do domínio em questão), respectivamente. Por conta do logaritmo na base 2, é medida na unidade de *bits*, e pode ser interpretada como a perda de informação ao analisar as informações referentes ao histograma como se fossem provenientes do processo gerador do histograma . De outra forma, podemos interpretá-la simplesmente como um “custo” para partir do histograma e “deformá-lo” até chegar no histograma , sendo essa relação não-simétrica. Fornecendo um exemplo musical, se e representam as frequências relativas dos tipos acordais nos *corpora* referentes aos compositores A e B, respectivamente, então mede a dissimilaridade entre os compositores A e B nesse aspecto, na direção de B para A. Após o cálculo das matrizes de dissimilaridades, aplica-se uma técnica de agrupamento hierárquico para cada domínio de análise, sendo o resultado *dendrogramas*, gráficos na forma de árvore que agrupam os *corpora* em diferentes níveis: no mais geral há somente um grupo com todos, e no nível mais baixo cada *corpus* compõe um único grupo; em níveis intermediários pode-se encontrar diversos padrões de similaridade, conforme está ilustrado na Figura 1. Por exemplo, o painel intermediário da Figura 1 ilustra o dendrograma dos *corpora* analisados com respeito à harmonia, e nota-se um agrupamento referente aos integrantes do movimento conhecido como “Tropicália” (composto pelos *corpora* de Rita Lee, Caetano Veloso e Gilberto Gil), bem como um grupo contendo somente Milton Nascimento, compatível com o distinto uso de modalismo em suas composições; o painel inferior da Figura 1 traz o dendrograma com respeito à relação entre melodia e harmonia, e nota-se o grupo contendo apenas Tom Jobim, que se ajusta ao fato de ser característico em sua música a presença marcante de tensões como notas estruturais. Tais análises mostram similaridades e agrupamentos que já são esperados, indicando que a técnica empregada pode ser uma poderosa ferramenta em estudos de *corpus*. Discussões mais aprofundadas com respeito a esse e outros resultados serão mostrados detalhadamente na apresentação.

**Palavras-chave**: Estudos de *corpus*. Música Popular Brasileira. Agrupamento hierárquico. Dendrograma. Teoria da Informação.