DOI: 10.5433/1981-8920.2018v23n3p314

PROCESSAMENTO DA LINGUAGEM NATURAL DO DOMÍNIO MUSICAL: DO SENTIDO À GESTÃO TERMINOLÓGICA NO AMBIENTE E-TERMOS

PROCESSING OF THE NATURAL LANGUAGE OF THE MUSICAL DOMAIN: FROM MEANING TO TERMINOLOGICAL MANAGEMENT IN THE ENVIRONMENT E-TERMOS

Juliana Rabelo do Carmo[a]
Valdirene Pereira da Conceição[b]

RESUMO
Introdução: A representação e recuperação da informação por meio do Processamento da Linguagem Natural (PLN) fundamenta a criação do vocabulário controlado de Música, com a finalidade de conhecer os processos de tratamento do conhecimento musical e extração de conceitos. **Objetivo:** O objetivo da pesquisa consiste em analisar o cenário prático-conceitual da indexação, visando a sistematização e organização de ferramentas de gestão terminológica e recuperação da informação por meio da estruturação de um vocabulário da área de Música. **Metodologia:** A pesquisa caracteriza-se como pesquisa aplicada, de natureza teórico-exploratória, utiliza os procedimentos de pesquisa bibliográfica e documental das áreas de Ciência da Informação, Linguística, Computação e Música. Emprega o modelo de pesquisa em PLN, com o *corpus* composto pela amostra de 10 dissertações e teses e 30 artigos de revistas científicas especializadas em Música, produzidos entre os anos de 2003 à 2013. Utiliza o ambiente colaborativo de gestão terminológica, E-termos para extração de termos. **Resultados:** Apresenta como resultados um corpora classificado como grande em termos quantitativos, fato que implica na obtenção de um nível de representatividade alta de termos, classificada como médio-grande para construção do vocabulário de Música. **Conclusões:** Finaliza indicando que o PLN se constitui de uma ferramenta efetiva para a tradução da linguagem natural, utilizando as expressões utilizadas para a busca da informação como objetos linguísticos.

[a] Mestranda em Ciência da Informação pela Universidade Federal de Santa Catarina (PPGCI-UFSC). E-mail: juliana.rabello@yahoo.com.br

[b] Doutora em Linguística e Língua Portuguesa pela Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP). Professora do Departamento de Biblioteconomia da Universidade Federal do Maranhão (UFMA). E-mail: cvaldirene@bol.com.br

Descritores: Organização da informação. Processamento da Linguagem Natural (PLN). Gestão terminológica. Vocabulário controlado de Música.

1 INTRODUÇÃO

O tratamento da informação em acervos especializados quando são derivados de uma precária análise documental revelam uma problemática encontrada para a organização e representação de conteúdos musicais, ocasionando assim a dificuldade na recuperação informacional. Esta questão se amplia quando se trata do aumento exponencial das fontes de informação em música torna a sua organização ainda mais complexa e revela uma realidade que carece de medidas direcionadas. Antonio (1994, p. 3) mostra que:

> [...] as dificuldades da comunidade musical na busca e sistematização das informações são crescentes [...] Essa situação aponta para a necessidade de desenvolver estudos que visem conhecer e sistematizar as condições da pesquisa e da organização da informação em música.

Diante de tal conjuntura, a necessidade de estudos nessa perspectiva possui uma longa trajetória de discussões na área da Ciência da Informação (CI), uma vez que os catálogos bibliográficos não compreendem a linguagem de indexação de modo a suprir as necessidades dos usuários na realização de buscas. Santini (2007, p. 12), explica que apesar de tais tentativas, os trabalhos científicos de Biblioteconomia relacionados com o tratamento e a Recuperação da Informação da Música (RIM) ainda são escassos, e conclui que:

> [...] as principais discussões de RIM são exploratórias e que as pesquisas em recuperação da informação da música estão em sua fase inicial. Muitas questões intrigantes permanecem sem investigação. Por exemplo, nenhum estudo rigoroso e compreensivo foi encontrado na literatura da área da Ciência da Informação [...].

Outro ponto que apoia a necessidade de organização e representação da informação musical consiste na grande produtividade de estudos científicos sobre Música nos principais centros de Pós-Graduação no Brasil nos últimos anos, fato percebido devido à vinculação de tais Programas à revistas científicas, Repositórios Institucionais e também na Biblioteca Digital de Teses e Dissertações (BDTD).

As motivações que levaram à sistematização da terminologia desta área do conhecimento, e o consequente interesse pela elaboração de um vocabulário são originados por diferentes ordens: profissional, na Ciência da informação, no sentido de identificar ferramentas de representação da informação musical, bem como da necessidade de instrumentos de controle terminológico nesta área que possibilite a identificação dos itens lexicais recorrentes.

O objetivo dessa pesquisa está em analisar o cenário prático-conceitual da indexação e representação da informação, visando a sistematização, organização de ferramentas de gestão terminológica e recuperação da informação por meio da estruturação de um vocabulário da área de Música.

A discussão proposta pretende, antes da atribuição de categorias e conceitos/léxicos para construção de um vocabulário, abordar o léxico musical, apontando para a interpretação correta dos termos tratados, situando-os na perspectiva da representação da informação, além de conferir-lhe o respectivo significado e possibilitando relações entre outros conceitos.

Portanto, este estudo visa beneficiar um universo variado de leitores sobre esta temática, como músicos, professores de música e de história da arte, amadores, estudantes de música, entre outros, indicando assim as contribuições interdisciplinares da Biblioteconomia através da aplicação de estudos sobre organização e representação da informação.

2 PROCESSAMENTO DA LINGUAGEM NATURAL: NÍVEIS E LIMITAÇÕES DE PLN

O PLN tem sido estudado pela área da CI na perspectiva teórica, em especial no campo da Indexação e Recuperação da Informação, por entender que os softwares baseados neste modelo propiciam a extração de termos com maior precisão semântica para recuperação da informação em sistemas de busca automatizados. McDonald e Yazdani (1990, p. 176), corroboram com a assertiva de que "[...] a pesquisa em PLN pode proporcionar *insights* bastante úteis sobre processos e representações da linguagem na mente humana, apontando, assim, para a verdadeira IA", por utilizar-se dos fundamentos linguísticos sobre os léxicos utilizados pelos usuários nos sistemas de buscas.

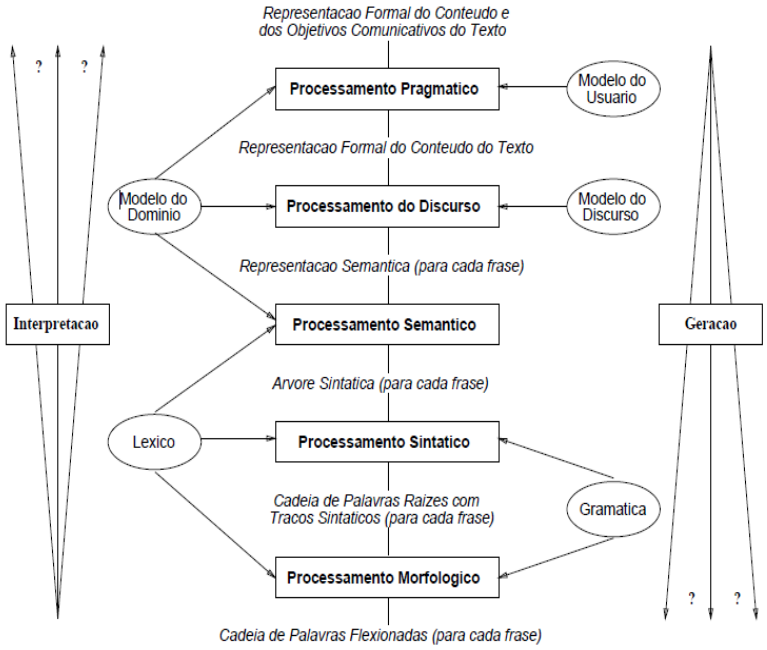
Em relação aos aspectos conceituais do PLN, também se percebe evoluções em relação às definições de alguns teóricos a partir de suas abordagens que contribuíram para constituição do objeto de estudo dessa área por meio de um encadeamento lógico.

O PLN subdivide-se em níveis de análise e/ou estudo que compreendem: a interpretação, onde são desenvolvidas questões relativas ao estudo da língua de modo que as palavras se tornem compreensíveis pelo computador e, consequentemente, o armazenamento para que ocorra a utilização destas palavras em sistemas, tomando como exemplo os tradutores (ou *chatterbots*); e de geração, que ocorre de forma inversa, a partir da inclusão de termos ou expressões, o computador adquire a capacidade de traduzir a compreensão do sistema para a linguagem natural por meio de estruturações semânticas pré-determinadas, no caso dos resumos e palavras-chave (CONTERATTO, 2006).

Tais estruturações fundamentam a arquitetura do PLN, apresentada por Nunes *et al.* (1999) e mostram que o banco de palavras, representado pelo Léxico, é acessado pelos analisadores Léxico, Sintático e Semântico, enquanto a Gramática serve ao analisador semântico para autenticar as palavras ou frases. Nesta perspectiva, o PLN enquanto Sistema baseado no Conhecimento

utiliza-se de cinco alicerces: gramática, léxico e o modelo de discurso, ou seja, as informações sobre a língua; modelo de domínio, a ser aplicado; e modelo do usuário que utiliza o sistema (NUNES *et al.,* 1999), ilustrados na Figura 1:

Figura 1 - Níveis de processamento em PLN.



Fonte: Nunes *et al.* (1999).

O nível morfológico consiste na definição da estrutura de palavras, bem como a significação e função de cada palavra na frase (adjetivo, substantivo, verbo, etc.); nível sintático, por meio da análise da construção gramatical, suas relações entre unidades linguísticas e sua colocação (sujeito, predicado verbal, etc.); nível semântico, onde as palavras são analisadas pelo seu significado, a partir da análise sintática; nível do discurso, compreensão do significado da palavra a partir do contexto em que ele está inserido; nível pragmático, onde ocorre a compreensão do conteúdo da frase ou texto, a partir da determinação de sua tipologia (pergunta, afirmação) (NUNES *et al.,* 1999).

Constituinte da principal dificuldade do PLN, a ambiguidade, ou seja, a pluralidade de sentidos de uma palavra tem sido uma das motivações para o aprimoramento dos modelos de aplicação do PLN, pois exige uma identificação

das unidades gramaticais, a serem aprofundadas nas próximas seções. Dentre as tipologias de ambiguidades encontradas para o PLN, podemos destacar de forma abreviada:

a) Homonímia lexical, quando uma mesma palavra possui significados que variam na escrita e no pronunciamento. Ex.: “Quem casa quer casa.”. A questão encontrada é, em qual dos dois momentos a palavra é verbo ou substantivo?
b) Ambiguidade sintática, quando a ambiguidade surge a partir da estruturação da palavra na frase. Ex.: “João pediu a Pedro para sair.”, onde a frase gera dois sentidos: 1) João está pedindo para Pedro autorização para sair, ou; 2) João está pedindo para Pedro se retirar.
c) Ambiguidade de finalidade, de forma resumida significa dúvidas no intuito da palavra no texto. Ex.: “Maria procura um banco.”. A indagação consiste em: o banco que Maria procura é um lugar para sentar ou um banco instituição financeira; neste caso, a solução seria o uso de um analisador de discurso.
d) Diferentes correferências possíveis, com abrangência em vários níveis linguísticos entre entidades. (RODRIGUES FILHO, 2004).

Nesse sentido, as soluções para esta problemática da ambiguidade estão indicadas no contexto de uso dos termos para assim apreender a sua significação. Com base nisto, revela-se a necessidade da análise linguística em diferentes níveis nas bases de conhecimento, por meio de abordagens que podem ser aliadas ao PLN, tomando como referenciais teóricos metodológicos aspectos morfossintáticos, semânticos e lexicais.

Presume-se que a utilização do PLN enquanto modelo computacional gera a observação de duas características de sua utilização, que diz respeito aos seus aspectos desfavoráveis e favoráveis, sendo que este último fornece uma abrangência maior do que as condições contrárias.

Algumas das limitações do PLN implicam no fato de possibilitar ao usuário a criação de suas próprias consultas, sem ter um padrão a seguir, o

que pode ocasionar resultados que não irão satisfazer as expectativas, além de erros na passagem da linguagem natural para uma linguagem de consulta, assim, o usuário pode criar uma resistência ao uso da aplicação ou, até mesmo, não acreditar nas respostas geradas pela mesma (SILVA; LIMA, 2007, p. 2).

Em contrapartida, dentre as vantagens do uso do PLN estão: a eliminação da necessidade de adaptação a formas inusitadas de interação, cuja construção gramatical costuma ser de difícil aprendizado e domínio, a exemplo das linguagens de consulta de bancos de dados (NUNES, 2007, *apud* NANTES, 2008, p. 26); o usuário não precisa entender o funcionamento de um banco de dados, ele apenas deseja que o resultado da pesquisa seja mostrado de forma simples e objetiva (GARIBA *et al.,* 2005 *apud* OLIVEIRA NETO; TONIN; PRIETCH, 2010, p. 2); é possível ainda, o entendimento de consulta com erros (termos digitados erroneamente) e incompletas, buscando por palavras próximas e pelo contexto da conversação (SILVA; LIMA, 2007, p. 2). Para tanto, basta que o usuário tenha um conhecimento básico da área - e ainda, assunto ou domínio -, da especialidade da base de dados.

3 O DOMÍNIO MUSICAL COMO MEIO DE EXPRESSÃO

A Música é um elemento fundamental nas diversas dimensões da vida humana. Como dado coletivo, social, a música reflete o meio ou a situação em que estamos inseridos. E neste mesmo sentido, marca e identifica gerações. Convém ainda ressaltar que o domínio musical é composto por informações de cunho artístico, que divergem da informação textual pela forma de expressão do seu conteúdo, que podem ser representados por sons, partituras, áudio digital, entre outros, que possuem uma linguagem que necessitam de representações específicas.

Na perspectiva filosófica educacional as influências das Artes, em especial a Música, tiveram seus reflexos registrados na história. Na Grécia, a Música era diretamente relacionada à Filosofia e à Educação, acreditava-se

que os seus efeitos agiam diretamente sobre a mente, corpo e alma, e por estes motivos, eram restritas somente aos cidadãos livres.

Granja (2006, p. 38) mostra a presença da Música nos conceitos do filósofo Platão na doutrina do *éthos musical*, ao representar que tais conceitos agregavam valores éticos e estéticos à alma, e explica:

> Os gregos chamavam de "*éthos musical*" o caráter particular associado a um determinado modo musical. Assim, um modo poderia exprimir um *éthos* do homem valente ou do homem sereno, enquanto outros estariam associados aos maus hábitos, à preguiça, à paixão (GRANJA, 2006, p. 38).

Trata-se de um elemento antropológico-cultural, que embora se origine e se desenvolva na esfera dos sentimentos, das emoções, do gosto pessoal, da sensibilidade e da subjetividade, tem também uma objetividade, cujas fontes se encontram no ambiente natural, histórico e social do povo. Para Queiroz (2000, p. 17),

> O conteúdo musical trata da mensagem, de caráter emocional, presente na música. Ou dizendo de outro modo, o conteúdo musical é aquilo que a música transmite, o estado que a música porta. E, por mais contestação que possa haver quanto à definição do quê ela porta, deve ser claro que algo ela porta. [...] quando verdadeiramente artístico, dá testemunho da verdade e da harmonia possível a vida – em uma forma compreensível à sensibilidade emocional.

Assim, pode-se afirmar que a Música reflete o que somos, o nosso modo de ser, de pensar as coisas, de relacionar-se com as pessoas e com o universo, ou seja, "[...] a música é uma arte eminentemente social, portanto, vinculada a sua época e ao seu lugar, suscetível às variações da sociedade incluindo evolução tecnológica" (FERREIRA, 2001, p. 92).

No campo da Antropologia, a Música é estudada como Etnomusicologia, que visa a inter-relação entre os aspectos históricos e culturais entre estas duas áreas, a exemplo disso, o antropólogo americano Alan P. Merriam (1964, p. 27), instituiu a teoria da Etnomusicologia e apregoa:

> A música é um fenômeno exclusivamente humano que só existe em termos de interação social; ela é feita por pessoas para outras pessoas, e isso é um comportamento aprendido. A música não existe por, de ou para si mesma. É preciso haver sempre os seres humanos a fazer algo para produzi-la. Em

> suma, a música não pode ser definida como um fenômeno do som sozinho, pois envolve o comportamento de indivíduos e grupos de indivíduos, e suas organizações particulares, exigências, e a concordância social das pessoas que a compõem.

O autor expõe que a Música constitui-se como um produto da cultura de um povo. Hummes ressalta dez funções da Música na Etnomusicologia, apresentada por Merriam, a saber: 1) expressão emocional; 2) prazer estético; 3) divertimento; 4) comunicação; 5) representação; 6) reação física; 7) impor conformidade às normas sociais; 8) validação das funções sociais e dos ritos religiosos; 9) contribuição para a continuidade e estabilidade da cultura e; 10) contribuição para a integração da sociedade (MERRIAM, 1964, *apud* HUMMES, 2004).

Estes elementos são resultados de processos e interpretações sociais, representados em diferentes formas. A Música é compreendida em seu contexto cultural e sua análise é feita de forma cognitiva. Nesse sentido, a preservação da Etnomusicologia mostra-se como elemento constituinte da memória, tradição e expressões artísticas de um lugar.

Por meio destas considerações, entende-se que a Música além de uma expressão artística, é entendida também como uma forma de linguagem, capaz de exprimir as realidades cotidianas, fato este que justifica as expressões musicais próprias de cada civilização ou povo, caracterizando-se como um instrumento de identidade e transformação social.

4 PERCURSO METODOLÓGICO

A compreensão e interpretação dos fenômenos a partir do contexto em que estão inseridos são fatores integrantes na produção do conhecimento. Para isso, é necessário o emprego de métodos para realização da pesquisa científica. Demo (1996, p. 34) aponta que a pesquisa é uma atividade cotidiana, considerando-a como uma atitude, um "[...] questionamento sistemático, crítico

e criativo, mais a intervenção competente na realidade, ou diálogo crítico permanente com a realidade em seu sentido teórico e prático."

O estudo trata-se de pesquisa aplicada (gerando aplicações práticas, dirigidas para problemas específicos), de natureza teórico-exploratória. Teórica, uma vez que investigamos em livros, artigos e afins a problemática da construção de vocabulários, genéricos e especializados, com base no tratamento computacional de dados linguísticos, seja por meio do PLN e suas ramificações, como: recuperação da informação, motores de busca, etiquetadores, desambiguadores, entre outros.

E ainda, de natureza exploratória ao propor de um vocabulário de um domínio de conceitos, com base no agrupamento de itens lexicais especializados, no caso, do campo musical. O percurso metodológico adotado para a realização da pesquisa norteia-se pelo procedimento de pesquisa bibliográfica e documental sobre as temáticas das áreas de Ciência da Informação, Linguística, Computação e Música, e por este motivo, assume o caráter de interdisciplinar, ao aplicar estes campos no modelo de pesquisa em Processamento Automático de Línguas Naturais (PLN).

4.1 O Ambiente e a Constituição do *Corpus* da Pesquisa

O *corpus* de Música envolveu etapas definidas, como levantamento de teses e dissertações referente ao domínio de Música, produzidas nos principais Programas de Pós Graduação em Música dos Centros/Instituições no Brasil, além de artigos científicos de revistas deste campo: Universidade de Campinas (UNICAMP); Universidade Federal do Rio Grande do Sul (UFRGS); Universidade Federal do Paraná (UFPR); Universidade Federal do Rio Grande do Norte (UFRN); Universidade Federal da Bahia (UFBA); Escola de Música (UFMG); Biblioteca Digital de Teses e Dissertações; Revista eletrônica de Musicologia; PERCEPTA – Revista de cognição musical; Música em perspectiva; Per Musi – Revista acadêmica de Música; Música e cultura –

Revista da Associação Brasileira de Etnomusicologia; Música em contexto; Revista Opus.

A definição da amostra compreende o período de 2003 à 2013, com *corpus* não estruturado, dentre os quais foram selecionados o quantitativo de 10 dissertações, 10 teses e em média 30 artigos - haja vista que a maioria das revistas possuem 2 publicações anuais -, produzidos por ano em todos os programas e/ou revistas citados. Vale ressaltar que tais documentos foram obtidos em formato eletrônico e em pdf, por meio de pesquisas na Web nos repositórios digitais dos Programas/Revistas citados.

A escolha dos *corpora* dos Programas de Pós-Graduação e Revistas citadas pautaram-se no fato de se tratar dos principais centros e veículos de comunicação científica do país voltados para a Música que disponibilizam suas produções em formato eletrônico em seus repositórios institucionais, se tornando assim fontes de literatura especializada. A seleção justifica-se ainda por tais Programas/Revistas se constituírem como fonte de coleta tanto dos conceitos quanto dos itens lexicais recorrentes no domínio de Música, contribuindo assim para solucionar a problemática da inexistência de instrumentos de controle terminológico para este domínio.

Os aspectos práticos da pesquisa foram executados diretamente pela plataforma E-termos, que oferece em um dos seus módulos a extração automática de unidades ou conjuntos lexicais que podem constituir uma unidade terminológica, ou seja, um termo, baseado em operações estatísticas proporcionadas pelo pacote *NSP statistic Pack-age*, com funções integradas no E-termos.

4.2 Extração Automática de Termos

A sistematização do vocabulário de Música na pesquisa abrange o uso do ambiente colaborativo de gestão terminológica, e-Termos, em especial, no que diz respeito a utilização de uma funcionalidade de extração de termos

deste ambiente, viabilizada pelo uso do s*oftware* estatístico, o Pacote *N-gram Stastistic Package* (NSP) em sua interface.

Desse modo, a extração automática a candidatos de termos tem com base o Processamento da Linguagem Natural, visando maior extração do conhecimento semântico dos textos processados. As etapas de fundamentação metodológicas são: a) Busca e seleção de fontes não estruturadas, no caso, teses e dissertações disponíveis em formato eletrônico e em pdf nas bases de dados definidas; b) Compilação do *corpus*: esta etapa envolve o armazenamento do *corpus*; c) Manipulação dos arquivos do *corpus*; d) Inclusão dos textos do *corpus* no e-Termos; e) Levantamento e análise da lista de unigramas, bigramas, trigramas, que correspondem a termos compostos por uma, duas ou três unidades, respectivamente, realizado pelo software Pacote NSP, integrado ao e-Termos; f) Limpeza das listas geradas, com eliminação de unidades que não correspondem a termos; g) Validação: escolha de julgadores, pertencentes a área abordada pelo; e preparação do material a ser conduzido para os julgadores para definição dos critérios para escolha dos termos definitivos; h) Identificação das categorias; i) Organização e apresentação do vocabulário de Música.

5 PROCESSAMENTO DA LINGUAGEM NATURAL DO DOMÍNIO MUSICAL

A abordagem utilizada é considerada semi-automática, devido à intervenção humana, que abarca quatro etapas principais: 1) construção de um vocabulário do Domínio Musical por meio da Indexação manual; 2) Compilação e processamento automático do corpus no ambiente E-termos; 3) Extração automática de candidatos a termos; e 4) Cálculo da frequência dos termos candidatos ao vocabulário, por meio de tarefas manuais e automáticas.

5.1 Sistematização Preliminar do Vocabulário Musical

A leitura técnica preliminar do *corpus,* em especial de informações contidas no título, palavras-chave, resumo, título das seções, introdução e

conclusão, torna possível a extração manual de candidatos a termos, com base na técnica da Indexação, haja vista que tais informações irão embasar o vocabulário do domínio musical. A motivação por este tipo de procedimento está, principalmente, no registro destas informações, entendidas aqui como categorias, ou seja, possíveis termos descritores, e itens lexicais, que posteriormente serão comparadas aos termos que serão extraídos automaticamente, visando analisar as aproximações dos resultados obtidos.

A estruturação do *corpus* foi feita a partir de etapas pré-definidas, como a busca em bases de dados de fontes não estruturadas específicas, para compilação (armazenamento) dos arquivos obtidos. Após esta etapa realizou-se a manipulação do *corpus*, que consiste na "limpeza" do texto nos formatos "doc", "HTML" e "pdf", transformando-os em formato "txt", também conhecido como bloco de notas, removendo todos os números, gráficos e/ou imagens do arquivo, tornando-o puramente textual com anotação e formatação, tornando o corpus mais manipulável para processamento computacional.

Após a indexação manual, fase que prepara a lista de termos em que a extração se baseia, e retoma-se o ideal de categorização a partir dos 699 textos, sendo estes, 358 dissertações, 100 teses e 241 artigos científicos indexados manualmente, que geraram o quantitativo de 29 categorias conceituais que abrigam as unidades lexicais que representam o Domínio Musical.

A elaboração de um corpus computadorizado obedece a critérios, preconizados por Sinclair (1991, a*pud* BERBER SARDINHA, 2000) como a representatividade, ou seja, a extensão do corpus que segundo o autor, deve ser o maior possível para obter representatividade. Para tanto, a título de informação, Berber Sardinha (2003) apresenta a classificação geral de tamanho de *corpus*, indicada na Tabela 1.

Tabela 1 - Classificação de tamanho de corpus

TAMANHO EM PALAVRAS	CLASSIFICAÇÃO
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio

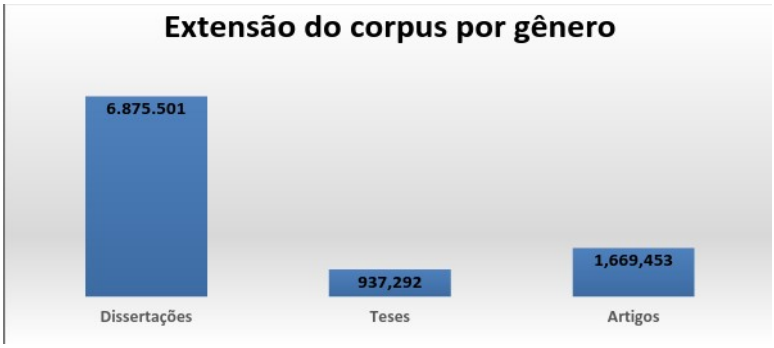
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Fonte: (BERBER SARDINHA, 2003).

Nessa perspectiva, ao final do processo compilatório, obteve-se um *corpus* médio-grande constituído por 9.482.246 palavras, extraídos de424 textos selecionados para o processamento de texto, extraídos de 14 fontes diferentes, entre produções técnico-científicas de Programas de Pós-Graduação e revistas científicas. Vale ressaltar que o quantitativo inicial de textos que participaram da indexação manual foi de 699 textos, com os cortes de 275 textos, tendo assim, como produto final 424 textos participantes.

Em termos quantitativos para fins de processamento automático pela plataforma E-termos, o corte citado justifica-se pela necessidade da exclusão dos textos que tiveram problemáticas tanto no momento da conversão do formato PDF para txt, para processamento do texto. Os documentos digitalizados também perpassaram pelo mesmo entrave, fato que, no caso do gênero Teses, justifica a redução de textos participantes no processamento de textos. Na etapa 2 do E-termos, a opção contador de palavras permite ao usuário medir o quantitativo de palavras do corpus analisado. Dividimos a extensão de acordo com o gênero técnico-científico adotado que compõe o *corpus* do domínio musical, conforme apresenta o Gráfico 1.

Gráfico 1 - Extensão do corpus por gênero



Fonte: dados da pesquisa.

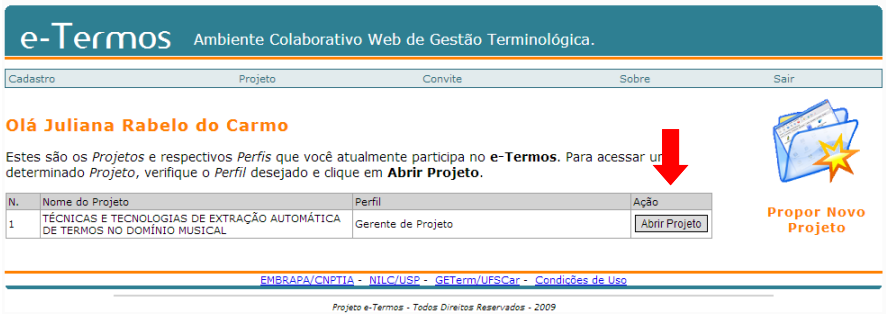
Desse modo, os dados revelam que o tamanho do corpus obtido sobre o domínio musical enquadra-se na classificação grande, por apresentar um total de 9.482.246 palavras. Na operacionalização do E-termos, a ferramenta

"Contador de frequência" da segunda etapa, apresenta ainda dados como total de palavras diferentes e índice de riqueza vocabular, e os resultados são:
a) Dissertações, com um total de palavras diferentes de 197.339, e índice de riqueza vocabular de 0.03; b) Teses, com 46.311 palavras diferentes e 0.049 de índice de riqueza vocabular, e; c) Revistas, com um total de 77.077 palavras diferentes e índice de riqueza vocabular de 0.046. Estas considerações indicam, além do quantitativo de palavras de cada corpus, o nível de representatividade e riqueza vocabular de termos, válidos para construção e análise do vocabulário controlado do domínio musical.

5.2 Vocabulário de Música: Estruturação e Análise

A pesquisa utilizou as ferramentas da plataforma colaborativa E-termos, apresentada em seções anteriores, que fornece acesso livre e gratuito. Para ter acesso é necessário a realização de um cadastro, e em seguida o pesquisador precisa propor um projeto para utilizar as ferramentas disponíveis. Após a aprovação da proposta, tem-se a opção de abrir o projeto para operacionalizar as etapas dispostas pelo ambiente, conforme a Figura 2.

Figura 2 - Página de identificação do projeto do usuário do E-termos.

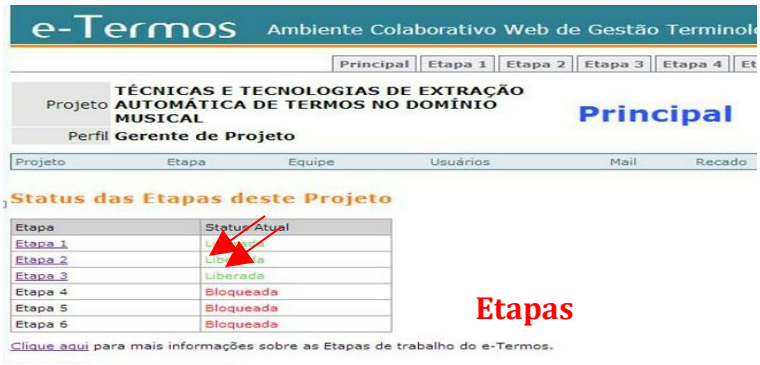


Fonte: https://www.etermos.cnptia.embrapa.br/main1.php.

Após clicar na ação de "Abrir Projeto", o usuário será direcionado para a página inicial, ilustrada pela Figura 3, de execução das etapas - já mencionadas anteriormente -, do E-termos, que para fins desta pesquisa

utilizaremos as etapas 2 que consiste no suporte e análise da qualidade dos corpus, e 3 que trata da extração automática de termos.

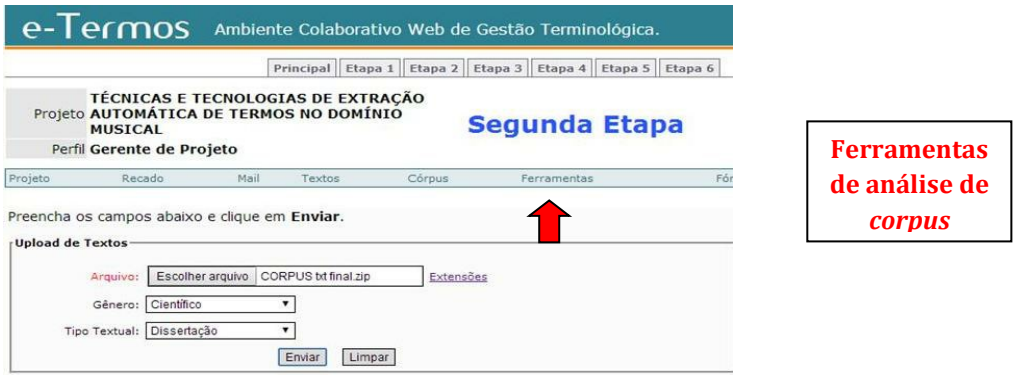
Figura 3 - Tela principal do E-termos



Fonte: https://www.etermos.cnptia.embrapa.br/main2.php.

Utilizamos a etapa 2 para fazer o upload dos textos que irão compor o *corpus* da pesquisa, que pode ser carregado de forma individual ou em pasta com extensão zipada, ou seja, os documentos ficam agrupados em um único arquivo. É necessário ainda estabelecer o gênero e o tipo textual dos textos que irão ser compilados para o corpus, como mostra a Figura 4.

Figura 4 - Segunda etapa do E-termos – Suporte e análise da qualidade dos córpus



Fonte: https://www.etermos.cnptia.embrapa.br/modulo2/modulo2.php.

A segunda etapa possibilita ainda como ferramentas da análise do *corpus*, o total de palavras no *corpus*, total de palavras diferentes e índice de riqueza vocabular. Na etapa 3 de extração automática de termos é realizada o

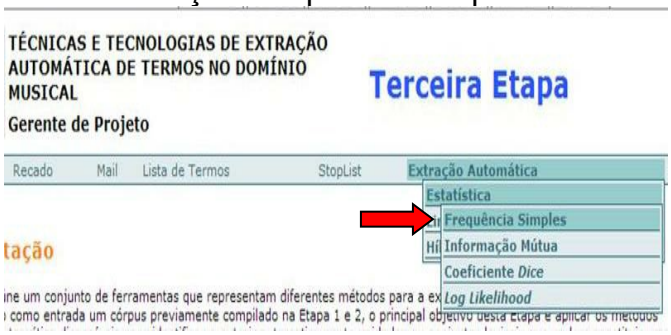
upload dos termos previamente estruturados por meio da indexação manual dos textos, conforme ilustra a Figura 5.

Figura 5 - Funções da aba “Lista de termos” na terceira etapa do E-termos

Fonte: https://www.etermos.cnptia.embrapa.br/modulo3/modulo3.php.

A figura 6 mostra a etapa seguinte, a extração de termos, que consiste, em outras palavras, na mineração de tais termos anteriormente inseridos no sistema, e sua verificação nos *corpus*, visando principalmente os índices de frequência simples, como fatores determinantes para as unidades ou conjuntos lexicais se caracterizarem como termos. Observa-se ainda, que o E-termos possibilita a extração Estatística, Linguística e Híbrida, porém, na prática somente a função Estatística estava disponível para uso.

Figura 6 - Terceira etapa do E-termos – Extração automática utilizando a função Frequência Simples



Fonte: https://www.etermos.cnptia.embrapa.br/modulo3/modulo3.php.

Na sequência, a tela apresentada na Figura 7, mostra as opções do extrator automático, onde será selecionado o *corpus* previamente inserido no

sistema na Etapa 2, o tamanho do termo que se deseja extrair do corpus, que pode ser de 1 a 7 n-grams, ou seja, de 1 a 7 unidades lexicais ou palavras.

Figura 7 - Opções do extrator na Terceira Etapa do E-termos

Fonte: https://www.etermos.cnptia.embrapa.br/modulo3/modulo3.php#.

As stoplists são as listas de palavras que não são necessariamente caracterizados como termos, mas que devem ser filtradas no momento do processamento da extração de termos, nesta questão, a plataforma dispõe da opção “Padrão do sistema”. A opção valor do corte inferior possibilita ao usuário estabelecer um valor de frequência mínimo para que uma palavra se torne candidata a termo. Para finalizar a fase de identificação que antecede a extração de termos, tem-se a identificação do resultado, ou seja, a nomeação da lista gerada após o processamento.

O sistema apresentou erro ao atender o comando de quatro a sete n-gram, e por este motivo, utilizamos os parâmetros de tamanho do termo (ou n-gram), unigrama (um), bigrama (dois) e trigrama (três) para execução da extração dos *corpus* de teses, dissertações e artigos. No que diz respeito a opção stoplist, a opção disponibilizada para uso foi “Padrão do sistema”.

O valor do corte inferior, ou seja, o quantitativo mínimo de frequência para que uma unidade lexical se candidate a termo, foi estruturada de acordo com a observação da frequência mínima de termos úteis, conforme apresenta a Tabela 2.

Tabela 2 - Valor dos cortes de frequência (termos desconsiderados) por gênero

GÊNERO	TAMANHO DO CÓRPUS	CORTE DE FREQUÊNCIA (QUANTITATIVO)
Teses	937.292	100 para unigramas 10 para bigramas e trigramas
Dissertações	6.875.501	100 para unigramas 10 para bigramas 10trigramas
Artigos científicos	1.669.453	100 para unigramas 10 para bigramas e trigramas

Fonte: dados da pesquisa.

Inicialmente foram observadas todas as frequências mínimas, e feita a limpeza manual de termos, para se estabelecer um quantitativo válido. Apesar de não haver um consenso na literatura da área sobre os valores de corte, Rijsbergen (1979, p. 21) afirma que "[...] uma certa arbitrariedade está envolvida na determinação dos pontos de corte, bem como na curva imaginária, os quais são estabelecidos por tentativa de erro." Nesse sentido, o corte de frequência da pesquisa se baseou nos parâmetros quantitativos de palavras que constituem os *corpus*, e ainda, de acordo com aqueles que mesmo com frequência baixa, no caso 10 (dez), se constituem como termos úteis.

Após a limpeza manual das listas de unigramas, bigramas e trigramas, por gênero textual, geradas pelo E-termos, e da eliminação das palavras que não se constituem necessariamente como um termo. Esse processo teve como resultado o comparativo entre os candidatos a termos por extração estatística e o número final de termos, resultantes da compilação de todos os gêneros textuais para uma visão geral dos dados de processamento de textos na Tabela 3.

Tabela 3 - Comparativo entre números de candidatos por extração estatística e número final de termos

N-gram	NÚMERO DE CANDIDATOS DO NSP (Dissertações, teses e artigos)	Nº FINAL DE TERMOS
Unigramas	4.880.851	930
Bigramas	673.644	513
Trigramas	927.961	226
Total	6.482.456	1.669

Fonte: dados da pesquisa.

Tais dados mostram que é possível afirmar que, neste caso, quanto maior o número de unidades que compõem o termo, maior o número de candidatos a termos, devido a função “Frequência simples” disponibilizada pelo pacote NSP, integrado ao E-termos. Outro fator que levou à grande redução de termos finais em relação aos candidatos extraídos pelo pacote NSP se deu pela grande quantidade de “sujeira” nos textos processados, a exemplo disso, podemos citar palavras com as acentuações que atrapalham o processo de extração e, consequentemente, descarta alguns termos que poderiam vir a ser úteis.

Apesar desse entrave, foi possível realizar a limpeza manual de algum destes termos, e visualizar a frequência de ocorrência, de acordo com o gênero, e o tamanho do n-gram desejado. Em suma, ao final do processo de extração automática de termos obteve-se, em geral, um quantitativo de 930 unigramas, 513 bigramas e 226 trigramas, dos três gêneros textuais analisados, totalizando 1230 termos considerados úteis e representativos para domínio musical. As verificações da coincidência entre as categorias que abrigam os termos, e obtidas por indexação manual e extração automática, estão agrupados na Tabela 4.

Tabela 4 - Comparativo entre categorias obtidas entre indexação automática e extração automática

CATEGORIAS DA INDEXAÇÃO MANUAL	CATEGORIAS DA EXTRAÇÃO AUTOMÁTICA
Análise musical	Análise musical
Arranjo	Arranjos
Aspectos emocionais e psicológicos da Música	Compositores
Compositores	Educação Musical
Educação musical	Ensino de Música
Escrita musical	Instrumento musical
Fisiologia vocal	Música vocal
Função social da Música	Nota musical
Gênero	Partitura
Grupo rítmico	Performance
História da Música	Prática musical
Instrumento musical	Processo composicional
Legislação	Processos de estúdio
Música e cultura	Uso da música
Música vocal	
Nota musical	
Performance	
Prática (campo) profissional	
Prática Interpretativa	

Processo composicional	
Processo de estúdio	
Produção vocal	
Psicologia cognitiva musical	
Recurso computacional	
Tempo musical	
Teoria musical	
Textura musical	
Uso (execução) da Música	
Vibração de instrumento	

Fonte: dados da pesquisa.

Dentre as categorias estruturadas, que abrigam os termos provenientes da extração automática, estão: análise musical, arranjos, compositores, educação musical, ensino de Música, instrumentos musicais, música vocal, nota musical, partitura, performance, prática musical, processos composicionais, processos de estúdio e uso da música.

Das 29 categorias manuais estruturadas, somente três indicaram baixa frequência de termos pertencentes na extração automática, a saber: fisiologia musical, função social da Música, grupos rítmicos, história da Música, legislação e textura musical. Tais considerações mostram que, apesar de o método estatístico gerar ruídos, ou seja, palavras que não possuem valor terminológico, este método é de extrema importância para fins de indexação, tradução, construção de tesauros, entre outras ferramentas de representação e recuperação da informação por proporcionar automatizar a identificação e seleção de unidades lexicais de um *corpus*, ao proporcionar rapidez na construção de Terminologias.

Os subsídios apontados a partir desta análise mostram que a extração automática de termos baseada em frequência estatística, facilitada pelo ambiente E-termos, permite o aprimoramento das técnicas e reforça a precisão no processo de Indexação. Espera-se ter demonstrado o percurso para estruturação de um vocabulário com os subsídios do Processamento da Linguagem Natural para construção de vocabulários.

6 CONCLUSÃO

A pesquisa se deteve em abordar a prática da indexação e representação da informação para estruturação do vocabulário do domínio musical no período de uma década, de 2003 a 2013, com a finalidade de analisar os léxicos produzidos nesta área e relacioná-los com outros conceitos, por meio da categorização.

Dentre os objetivos da análise prático-conceitual da indexação, destacam-se como resultados os instrumentos de representação da informação musical, como catálogos, índices e tesauros com o objetivo de servir como recurso e/ou ferramenta informacional auxiliar para pesquisadores e interessados em geral na busca e recuperação da informação musical. Porém, percebe-se que os instrumentos existentes ainda são escassos e falhos no que diz respeito ao conteúdo dos documentos, para o suprimento da necessidade informacional de seus usuários.

Em se tratando da área de Música, não foi localizada obra lexicográfica e/ou terminológica sobre o domínio, com contribuições dos aportes automáticos, em especial de PLN, para sua estruturação, o que evidencia a contribuição desta pesquisa ao tentar minimizar a carência de informações sobre o controle do léxico utilizado pela área musical.

Observa-se ainda a importância da categorização para a identificação de assuntos de um domínio de especialidade, devido à possibilidade de abordagem facetada de conteúdos que permite a visualização de uma área do conhecimento como um todo sistematizado, e viabilizando relações com outros conceitos.

O resultado deste processo foi a atribuição de 29 categorias, sendo que destas, as que tiveram mais termos agregados na Indexação Manual foram Prática interpretativa (121 termos), Processo composicional (100 termos) e Educação Musical (97 termos). A extração automática por meio do E-termos, por sua vez, mostrou-se eficaz ao apresentar um quantitativo de 922 itens

lexicais, entre unigramas, bigramas e trigramas que podem aprimorar a construção de vocabulários fundamentados em PLN.

Com relação à análise comparativa entre os processos manual e automático de extração de termos, a convergência entre estes dois métodos consiste na subjetividade humana para seleção e correção dos termos encontrados, porém, vale ressaltar que a intersecção entre as categorias obtidas por Indexação manual e as categorias geradas por extração automática alcançaram índices de frequências diferentes durante o processo. Isso significa que, nem todos os termos elencados no método manual foram extraídos automaticamente.

Nesse sentido, a escolha pelo método estatístico se deu pelo fato de a frequência apresentar um quantitativo maior de descritores significativos, que proporcionam uma representação e recuperação mais precisa de termos, além da rapidez na extração diante de grandes volumes de textos.

A principal problemática encontrada durante o processo de extração automática de candidatos a termos consiste na disponibilidade dos softwares (híbridos, linguísticos ou estatísticos) de forma gratuita, haja vista que tais softwares citados durante o referencial teórico encontram-se em teste nos Programas de Pós-Graduação voltados para a Computação/Inteligência Artificial e Linguística Computacional e, por este motivo, ainda não foram disponibilizados para a comunidade acadêmica.

Outra questão envolve a interface com o usuário, a exemplo do *GATE* e Pacote NSP que foram idealizados inicialmente para a pesquisa, porém, os seus formatos via linha de código, ou seja, onde o software para seu funcionamento necessita de códigos para programação de computadores, ocasionaram entraves não somente para este estudo.

Vale retomar algumas considerações a respeito da extração automática, e suas abordagens observadas no desenvolvimento da pesquisa: na Terminologia, a extração automática corresponde à aquisição de um produto terminológico que representa os léxicos, a exemplo de dicionários, índices ou glossários; enquanto a Computação a entende como abordagem automática de

reconhecimento e extração de termos de uma especialidade, geralmente realizada por meio das ferramentas de PLN.

A extração de candidatos a termos se mostra proveitoso também para outras utilidades, tais como: recuperação da informação, construção de ontologias, sumarização automática, tradutores, alinhamento de textos, corretores gramaticais e PLN. Com a utilidade da extração automática minimiza-se a subjetividade do indexador por meio do uso dos subsídios da Inteligência Artificial que tende a aproximação do raciocínio homem-máquina para solução de questões relacionadas à linguagem.

Apesar de todo o vasto subsídio teórico disponível pela Linguística Computacional, em especial pelo PLN, em termos práticos ainda existem lacunas na prática no Brasil, que se encontra ainda em fase de desenvolvimento e aprimoramento de softwares de forma a conduzi-los a efetivação da interpretação da linguagem humana. O que nos leva a crer que ainda existe um longo percurso a ser feito para abranger todos os domínios do conhecimento.

A literatura revela que grande parte das pesquisas em PLN concentra-se em áreas e subdomínios: manipulação de bases de dados, sistemas tutores, sistemas de processamento de textos científicos, sistemas especializados, tradução automática, sistemas acadêmicos, geração de resumos e extração da informação.

Assim com base nos pressupostos teórico-metodológicos dos estudos do Léxico, Linguística de *Corpus* e PLN estudados e empregados, foi possível a elaboração do vocabulário de Música, a ser estendido e aprimorado. Embora se tenham atingido os propósitos aqui pretendidos, é importante considerar os passos previstos para continuidade da pesquisa, como:

- Extração de termos baseada em técnicas estatísticas, para execução de outras ferramentas;
- Criação e processamento de um novo corpus que envolva outras tipologias de produções técnico-científicas do domínio musical;
- Construção de Ontologias do domínio musical.

O PLN mostrou-se como uma ferramenta eficaz para processamento de grandes volumes de dados, com muito a contribuir no que diz respeito à redução do tempo de desempenho de tarefas de mineração de textos e ao possibilitar a identificação dos termos mais utilizados para representação de um domínio. Apesar destas contribuições, destaca-se que a intervenção humana ainda é necessária para a limpeza dos materiais obtidos e para a validação dos resultados.

Espera-se ter proposto uma metodologia para elaboração de novas representações de domínios, com o intuito de aprimorar as técnicas e ferramentas de representação da informação utilizadas na Ciência da Informação.

Diante de tais considerações chega-se a conclusão de que o vocabulário de Música é uma ferramenta útil e enriquecedora no sentido da apresentação tanto para os próprios músicos, ao facilitar a aproximação e a consequente recuperação dos termos utilizados na linguagem natural para a linguagem artificial no momento de busca em sistemas, por meio do PLN, quanto para os linguistas e bibliotecários, ao fornecer subsídios que contribuem para o controle dos léxicos.

REFERÊNCIAS

ANTONIO, I. **Informação e Música no Brasil:** memória, história e poder. São Paulo, 1994. 285 f. Dissertação (Mestrado em Ciência da Informação) – Universidade de São Paulo, Escola de Comunicações e Artes, 1994.

BERBER SARDINHA, T. **O que é um corpus representativo.** São Paulo: LAEL PUCSP, 2000.

_____. Tamanho de corpus. **The Especialist**, São Paulo, v. 23, n. 2, p. 103-122, 2003.

CONTERATTO, G. B. H. Semântica e computação: uma interação necessária para o aperfeiçoamento de sistemas PLN. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 353-367, jun. 2006.

DEMO, P. **Pesquisa e construção de conhecimento.** Rio de Janeiro: Tempo Brasileiro, 1996.

FERREIRA, S. **O ensino das artes:** construindo caminhos. Campinas: Papirus, 2001.

GRANJA, C. E. de S. C. **Musicalizando a escola:** música, conhecimento e educação. São Paulo: Escrituras Ed., 2006.

HUMMES, J. **As funções do ensino de música, sob a ótica da direção escolar: um estudo nas escolas de Montenegro/RS.** Dissertação (Mestrado em Educação Musical) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

McDONALD, C.; YAZDANI, M. **Prolog programming:** a tutorial introduction. Oxford: Blackwell Scientific Publications, 1990.

MERRIAM, A. P. **The anthropology of music.** Evanston: Northwestern University Press, 1964.

NANTES, L. M. **Desenvolvimento de um sistema baseado em linguagem natural para consultas em banco de dados na Web.** 2008. 63 f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade do Oeste Paulista, Presidente Prudente, 2008. Disponível em: <http://fipp.unoeste.br/~chico/FIPP/projetos/projeto2008/Monografia_Nantes_2008.pdf>. Acesso em: 20 ago. 2013.

NUNES, M. G. V.; *et al.* **Introdução ao processamento das línguas naturais**. Notas Didáticas do ICMC, n. 38. São Carlos/SP, 1999.

OLIVEIRA NETO, J. M.; TONIN, S. D.; PRIETCH, S. S. **Processamento de linguagem natural e suas aplicações computacionais.** 2010. Disponível em: <http://www.inpa.gov.br/erin2010/Artigo/Artigo9.pdf>. Acesso em: 20 ago. 2013.

RODRIGUES FILHO, I. W. **Processamento de linguagem natural.** 2004. Disponível em: <http://www.inf.ufsc.br/~ilson/slides.ppt>. Acesso em: 15 mar. 2014.

QUEIROZ, G. J. P. de. **A música compõe o homem, o homem compõe a música.** São Paulo: Cultrix, 2000.

RIJSBERGEN, V. C. J. **Information retrieval.** 2. ed. Glasgow: Dept. of Computer Science, University of Glasgow, 1979.

SANTINI, R. M. Recuperação da informação de música e a Ciência da Informação: tendências e desafios de pesquisa. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 8, 2007, Salvador. **Anais...** Salvador: UFBA, 2007. p. 1-14.

SILVA, R. R.; LIMA, S. M. B. Consultas em bancos de dados utilizando linguagem natural. **Revista Eletrônica da Faculdade Metodista Granbery**, Juiz de Fora, v. 7, n. 2, ago./dez. 2007. Disponível em: <http://re.granbery.edu.br/artigos/MjQ0.pdf>. Acesso em: 30 ago. 2013.

PROCESSING OF THE NATURAL LANGUAGE OF THE MUSICAL DOMAIN: FROM MEANING TO TERMINOLOGICAL MANAGEMENT IN THE ENVIRONMENT E-TERMOS

ABSTRACT
Introduction: The representation and retrieval of information through Natural Language Processing (PLN) is based on the creation of the controlled vocabulary of Music, with the purpose of knowing the processes of musical knowledge and extraction of concepts. **Objective:** The objective of the research is to analyze the conceptual and conceptual scenario of indexation, aiming at the systematization and organization of terminological management tools and information retrieval through the structuring of a vocabulary in the Music area. **Methodology:** The research is characterized as applied research, of a theoretical-exploratory nature, using the bibliographic and documentary research procedures of the areas of Information Science, Linguistics, Computing and Music. It uses the PLN research model, with the corpus composed by the sample of 10 dissertations and theses and 30 articles of specialized scientific journals in Music, produced between the years 2003 and 2013. It uses the collaborative environment of terminological management, E-terms for extraction of terms. **Results:** It presents as results a corpora classified as great in quantitative terms, fact that implies in obtaining a high level of representativeness of terms, classified as medium-large for the construction of Music vocabulary. **Conclusions:** It ends by indicating that PLN is an effective tool for the translation of natural language, using the expressions used to search for information as linguistic objects.

Descriptors: Organization of information. Natural Language Processing (PLN). Terminological management. Vocabulary of Music.

PROCESAMIENTO DEL LENGUA NATURAL DEL DOMINIO MUSICAL: DEL SENTIDO A LA GESTIÓN TERMINOLÓGICA EN EL MEDIO AMBIENTE E-TÉRMINOS

RESUMEN
Introducción: La representación y recuperación de la información a través del Procesamiento del Lenguaje Natural (PLN) fundamenta la creación del vocabulario

controlado de Música, con la finalidad de conocer los procesos de tratamiento del conocimiento musical y la extracción de conceptos. **Objetivo:** El objetivo de la investigación consiste en analizar el escenario práctico-conceptual de la indexación, visando la sistematización y organización de herramientas de gestión terminológica y recuperación de la información por medio de la estructuración de un vocabulario del área de Música. **Metodología:** La investigación se caracteriza como investigación aplicada, de naturaleza teórico-exploratoria, utiliza los procedimientos de investigación bibliográfica y documental de las áreas de Ciencia de la Información, Lingüística, Computación y Música. Se emplea el modelo de investigación en PLN, con el corpus compuesto por la muestra de 10 disertaciones y tesis y 30 artículos de revistas científicas especializadas en Música, producidos entre los años 2003 a 2013. Utiliza el ambiente colaborativo de gestión terminológica, E-términos para extracción de términos. **Resultados:** Presenta como resultados un corpora clasificado como grande en términos cuantitativos, hecho que implica la obtención de un nivel de representatividad alta de términos, clasificada como medio-grande para la construcción del vocabulario de Música. **Conclusiones:** Finaliza indicando que el PLN se constituye de una herramienta efectiva para la traducción del lenguaje natural, utilizando las expresiones utilizadas para la búsqueda de la información como objetos lingüísticos.

Descriptores: Organización de la información. Procesamiento del Lenguaje Natural (PLN). Gestión terminológica. Vocabulario controlado de Música.