DOI: 10.5433/2237-4876.2023v26n3p102-114

O uso de técnicas computacionais no âmbito da Dialetologia e da Lexicografia: XML e X-Query

Jorge Luiz Nunes dos **SANTOS JUNIOR***

* Doutor em Letras pela Universidade Federal de Mato Grosso do Sul (UMFS) e pós-doutorando na Universidade Federal da Grande Dourados (UFGD/CAPES). E-mail: jorgesantosjunior@gmail.com.

Resumo:
Esse trabalho tem como objetivo demonstrar e discutir sobre o uso da *Extensible Markup Language* (*XML)* e das expressões *X-Query* como técnicas computacionais utilizadas no tratamento de dados dialetais e lexicográficos permitindo, dessa forma, a execução de tarefas de Processamento de Linguagem Natural (PLN). Trata-se de um recorte da pesquisa de doutoramento que teve como objetivo, mais amplo, desenvolver um protótipo de vocabulário dialetal eletrônico a partir dos dados do Projeto Atlas Linguístico do Brasil (ALiB), referente à rede de pontos do interior da região Norte do país. Desse modo, foi possível recuperar informações específicas do banco de dados em *XML* para analisar questões de interesse da Dialetologia, a partir da filtragem de dados por meio das variáveis localidade, sexo e idade, bem como selecionar um conjunto de informações em formato lexicográfico para exibi-las no protótipo do vocabulário dialetal. O estudo fundamenta-se na Linguística Computacional, na Dialetologia e na Lexicografia e justifica-se pela necessidade da transformação de dados de origem oral em um formato que permita o PLN, viabilizando a recuperação eletrônica de informações para análises linguísticas, além de fornecer meios de apresentação de dados em formato digital.

Palavras-chave:
Processamento de Linguagem Natural. Dialetologia. Lexicografia.

Signum: Estudos da Linguagem, Londrina, v.26, n.3, p.102-114, dezembro. 2023
Recebido em: 19/09/23
Aceito em: 14/03/24

O uso de técnicas computacionais no âmbito da Dialetologia e da Lexicografia: *XML* e *X-Query*

Jorge Luiz Nunes dos Santos Junior

INTRODUÇÃO

Ao refletir sobre as metodologias utilizadas no âmbito dos estudos linguísticos e, mais especificamente, ao se focalizar os projetos desenvolvidos no campo da Dialetologia e da Lexicografia, é possível constatar que a metodologia da Linguística de *Corpus* tem fundamentado pesquisas nessas disciplinas, tendo em vista a inerente necessidade de se trabalhar com um grande volume de dados. Nesse cenário, um *corpus* é compreendido como a sistematização criteriosa de uma grande quantidade de textos devidamente selecionados com a finalidade de servir aos propósitos de um dado estudo científico (Berber-Sardinha, 2004, p. 4).

Destaca-se que os estudos baseados em *corpus* são antigos e, de acordo com O'Keeffe e McCarthy (2010, p. 3), estudiosos da Bíblia Sagrada já produziam, séculos atrás, listas de palavras com a anotação da frequência e do contexto de uso. Isso revela que a Linguística de *Corpus* é anterior à era digital.

Todavia, foi com o advento do computador – em meados dos anos 1940 – e, posteriormente, com o desenvolvimento de softwares de processamento e análise de texto, que a Linguística de *Corpus* se automatizou e ganhou um novo panorama no universo científico. Assim, as listas de palavras com dados estatísticos que eram produzidas manualmente no passado são, atualmente, geradas por computadores.

Nesse panorama, o surgimento de programas de computador como *WordSmith Tools*[1], *AntConc*[2], *Sketch Engine*[3], *LancsBox*[4] e *FieldWorks*[5] têm proporcionado aos linguistas a oportunidade de ampliarem as perspectivas de estudos a partir do processamento eletrônico de dados lexicais.

Porém, mesmo com a inovação tecnológica desses recursos computacionais, as funcionalidades desses softwares são limitadas, pois são capazes de executar apenas as tarefas de Processamento de Linguagem Natural (PLN) para as quais foram projetados. Dessa forma, caso o pesquisador necessite realizar diferentes etapas de PLN, não poderá contar com o uso desses programas, pois não é possível adicionar novas funcionalidades aos softwares disponíveis no mercado.

Dessa maneira, nos casos em que uma pesquisa científica necessite ir além da execução das tarefas para as quais os programas de computador foram projetados, o pesquisador poderá recorrer a duas alternativas, a saber: i) contratar um especialista em programação para desenvolver o

[1] Software pago desenvolvido por Mike Scott, em meados de 1996. Para mais informações acesse: https://lexically.net/wordsmith/. Acesso em: 15 set. 2023.

[2] Software gratuito desenvolvido por Anthony Laurence, por volta de 2002. Para mais informações acesse: https://www.laurenceanthony.net/software/antconc/. Acesso em: 15 set. 2023.

[3] Plataforma on-line de acesso pago fundada por Adam Kilgarriff, em 2003. Para mais informações acesse: https://www.sketchengine.eu/. Acesso em: 15 set. 2023.

[4] Software gratuito desenvolvido por Vaclav Brezina, William Platt e Tony McEnery, em meados de 2015. Para mais informações acesse: http://corpora.lancs.ac.uk/lancsbox/. Acesso em: 15 set. 2023.

[5] Software para a produção de obras lexicográficas desenvolvido pelo grupo *SIL International.* Para mais informações acesse: https://software.sil.org/fieldworks/.

software de que precisa; ii) adquirir os conhecimentos necessários para construir suas próprias ferramentas computacionais.

A primeira opção sugerida, embora demande investimento financeiro, se configura em uma alternativa mais rápida, caso o pesquisador não disponha de tempo para aprender o essencial para desenvolver seu próprio software. Todavia, se o estudioso tiver alguns meses para se capacitar, além de uma orientação especializada que o norteie sobre quais conteúdos do campo da Ciência da Computação deverá compreender, a segunda opção, certamente, será a escolha que mais trará resultados ao estudioso.

É necessário ressaltar que desenvolver as próprias ferramentas computacionais a fim de atender aos objetivos de uma pesquisa científica não significa dominar uma série de temas relacionados à Ciência da Computação, mas apenas aqueles conteúdos considerados essenciais para atender aos objetivos do projeto. Dessa forma, a assessoria de um especialista do ramo computacional se faz importante no sentido de guiar o pesquisador pelo vasto mundo da Informática, pois somente esse profissional saberá o melhor caminho a seguir, tendo em vista que uma ferramenta computacional pode ser desenvolvida de maneiras diferentes.

Desse modo, o uso de metodologias computacionais no âmbito dos estudos dialetais e lexicográficos fornecem subsídios para a estruturação de dados em formato eletrônico que podem ser manipulados para atender aos propósitos dessas áreas, além de permitir o desenvolvimento de produtos digitais. Para tanto, o objetivo deste artigo é apresentar e discutir como a construção de um banco de dados em *Extensible Markup Language* (*XML)* foi útil para a organização de dados orais de maneira sistematizada e, a partir da escrita de expressões *X-Query*, demostrar como foi possível recuperar informações dialetais específicas em um ambiente informatizado e exibi-las em formato lexicográfico.

Destaca-se que as técnicas computacionais aplicadas à Dialetologia e à Lexicografia, demostradas e discutidas neste artigo, representam um recorte de alguns dos resultados alcançados durante a pesquisa de doutoramento desenvolvida no âmbito do Programa de Pós-Graduação em Letras da Universidade Federal de Mato Grosso do Sul, *campus* Três Lagoas (UFMS/CPTL), defendida em 2023, que teve como objetivo, mais amplo, a criação de um protótipo de vocabulário dialetal eletrônico a partir dos dados do Projeto Atlas Linguístico do Brasil (ALiB), referente à rede de pontos do interior da região Norte do país.

Assim, no que diz respeito à Dialetologia conseguiu-se organizar os dados de origem oral de modo a ser possível a recuperação do conteúdo das entrevistas a partir de filtros que selecionam dados por meio das variáveis presentes no *corpus* do ALiB, ou seja, filtrar as respostas das entrevistas pelas variáveis localidade, sexo, idade e escolaridade.

No que concerne à Lexicografia, os elementos que formam a microestrutura lexicográfica do protótipo do Vocabulário Dialetal do interior da região Norte (VoDiNorte) foram estruturados no banco de dados em *XML,* de modo a permitir que a recuperação de dados específicos da microestrutura funcionasse em conjunto com uma página web[6] que foi desenvolvida, posteriormente, para exibir as informações dialetais que receberam o tratamento lexicográfico.

Desse modo, uma interdisciplinaridade entre Dialetologia, Lexicografia e técnicas específicas da Ciência da Computação formam a base teórico-metodológica empregada na pesquisa de Doutorado que proporcionou uma ampliação nas perspectivas de estudos, bem como o desenvolvimento de ferramentas computacionais de análises lexicais e de apresentação de dados.

Ressalta-se, ainda, que este estudo pretende não apenas expor uma metodologia computacional para processar e apresentar dados dialetais em formato lexicográfico, como também visa a convidar pesquisadores da área de Letras a considerarem as possibilidades de aplicarem o *XML* e as expressões *X-Query* em seus projetos, tendo em vista que o investimento necessário para o implemento dessas tecnologias no âmbito da pesquisa científica pode ser feito com treinamentos e/ou disciplinas específicas ministradas em cursos de pós-graduação.

[6] Por se tratar de um protótipo, a página web do VoDiNorte encontra-se indisponível ao público.

1. FUNDAMENTAÇÃO

No decorrer do desenvolvimento dos estudos linguísticos é possível identificar que, o Distribucionalismo de Bloomfield (1933) e o Gerativismo de Chomsky (1965) apresentaram propostas de análise da linguagem por meio de um paradigma matemático que, de certa forma, influenciaram o desenvolvimento de técnicas computacionais para o processamento de dados linguísticos.

Assim, observa-se que o Distribucionalismo defende um método de estudo mecanicista em que a estatística é um dos elementos utilizados para quantificar e descrever as regularidades identificadas na língua. Destaca-se, ainda, que a abordagem de Bloomfield exclui questões relacionadas à história e à semântica, defendendo apenas uma análise da distribuição dos elementos linguísticos dispostos na organização dos enunciados produzidos pelos falantes (Orlandi, 2009, p. 31-32).

O Gerativismo, por sua vez, fez contribuições importantes ao estabelecer um conjunto de regras com a finalidade de descrever o funcionamento da Gramática Gerativa. Assim, a partir da análise dos elementos linguísticos que se combinam em uma sentença, que correspondem aos níveis fonológico, sintático e semântico (Chomsky, 1965, p. 15), é possível identificar padrões e quantificar a posição em que cada um dos elementos linguísticos aparecem no contexto de uso de milhares de textos, para prever as possibilidades de combinações aceitas pela gramática de uma dada língua natural.

Desse modo, as contribuições de Bloomfield (1933) e Chomsky (1965) permitem observar o funcionamento das línguas naturais por meio de um sistema de regras e de probabilidades. Essa abordagem se aperfeiçoou ao longo dos anos e serviu de inspiração para os estudos no campo da Linguística Computacional como, por exemplo, o desenvolvimento das ferramentas de geração automática de texto como é o caso dos corretores automáticos, que fornecem sugestões ortográficas ao usuário a partir do acesso a um dicionário, além de sugestões sintáticas por meio de dados estatísticos.

Destaca-se, ainda, que esse método estatístico de processamento das combinações morfossintáticas das línguas naturais evoluiu de tal modo que, atualmente, as técnicas de PLN de última geração permitem que o homem manipule dados textuais por meio dos computadores a partir do uso das linguagens de programação e, mais recentemente, com desenvolvimento de tecnologias baseadas em Inteligência Artificial como o *ChatGPT,* é possível realizar operações de PLN sem a necessidade de dominar as linguagens de programação. Assim, o pesquisador pode descrever o tipo de PLN que necessita desenvolver para um determinado projeto que a máquina gera, a partir de orientações específicas, as linhas de código que deverão ser executadas em um ambiente informatizado. Desse modo, o *ChatGPT* pode ser visto como um operário que realiza o serviço pesado em uma construção, enquanto o pesquisador, nessa analogia, pode ser aludido ao supervisor da obra, sendo o responsável por gerenciar todas as etapas da empreitada.

Em relação à Dialetologia, o desenvolvimento de ferramentas para o processamento de dados orais apresenta singularidades importantes e a distinção do tipo de informante e das localidades em que as entrevistas foram realizadas são pontos fundamentais nos estudos dialetais (Chambers; Trudgill, 1994, p. 87-88). Desse modo, as ferramentas de PLN devem ser capazes de filtrar e exibir informações a partir das características do *corpus* oral em questão e, no caso dos dados coletados pelo Projeto ALiB, significa permitir a manipulação de informações por meio da seleção das variáveis localidade, sexo, idade e escolaridade. Para tanto, é preciso que o pesquisador utilize a linguagem de marcação *XML* para delimitar essas variáveis dialetais durante o processo de transcrição. Na prática, isso significa armazenar cada especificidade do *corpus* em determinados campos do arquivo *XML* permitindo, dessa forma, a recuperação eletrônica dos dados a partir dos critérios especificados na estrutura do *XML* (Walmsley, 2015, p. 1-2).

Deve-se ressaltar que, o banco de dados sem um mecanismo de busca não é capaz de exibir

resultados ao usuário. Dessa forma, se faz necessário o desenvolvimento de ferramentas computacionais que funcionem como motores de pesquisa e que possam ser criadas a partir da linguagem de consulta *X-Query* que, segundo Walmsley (2015, p. 2) possibilita a busca de informações textuais armazenadas em formato *XML*, exibindo-as de modo estilizado, além de viabilizar a construção de aplicações web complexas.

Vale acrescentar que esse princípio metodológico também se aplica à Lexicografia, bem como a qualquer outra área do conhecimento, pois a linguagem de marcação *XML* permite estruturar qualquer tipo de dado com a finalidade de executar a recuperação eletrônica de informações a partir das expressões *X-Query*. Nesse sentido, a construção de um banco de dados em *XML* pode assumir formas estruturais distintas que corresponderão ao tipo de informação que o pesquisador desejará processar.

Em síntese, as técnicas de PLN aplicadas aos estudos dialetais se inserem no campo da Linguística Computacional, que abrange tanto o uso de softwares existentes no mercado para o estudo da linguagem, quanto o desenvolvimento de ferramentas computacionais personalizadas para a manipulação digital de dados linguísticos (Srinivasa-Desikan, 2018, p. 11).

Outro ponto que deve ser mencionado é o impacto do uso de bancos de dados e de motores de busca no desenvolvimento de produtos lexicográficos. Nessa perspectiva, a incorporação dessas tecnologias ao labor dicionarístico permite o planejamento e a execução de projetos que funcionam como plataformas on-line de exibição de dados tradados lexicograficamente. Esse é o novo paradigma da Lexicografia Eletrônica que está em curso, ou seja, desenvolver produtos lexicográficos fazendo uso, desde o início do projeto, de tecnologias disruptivas (Tarp, 2019, p. 255).

2. METODOLOGIA

A metodologia empregada na pesquisa de doutoramento pode ser utilizada em diferentes tipos de projetos. Neste artigo em específico, busca-se apresentar um recorte das técnicas empregadas na Tese (Santos Junior, 2023), incluindo os procedimentos necessários para se realizar tarefas de PLN a partir do banco de dados em *XML*. Assim, é possível sintetizar o processo metodológico em dois blocos, a saber: i) a construção do banco de dados em *XML* e o armazenamento dos dados; ii) o desenvolvimento das ferramentas computacionais de PLN.

Os dados de origem orais utilizados na Tese de Doutorado e, por sua vez, demonstrados neste artigo, são originários do Projeto ALiB[7] que se caracteriza por um trabalho interinstitucional, com sede na Universidade Federal da Bahia (UFBA), que tem como objetivo principal descrever a língua portuguesa do Brasil por meio de uma perspectiva diatópica[8].

Dessa forma, o banco de dados em *XML* foi projetado para armazenar[9] todas as informações contidas nas entrevistas do Questionário Semântico-lexical realizadas pelo ALiB (QSL/ALiB) de maneira a ser possível solicitar ao sistema a recuperação de dados por meio de filtros. Como o objetivo da Tese, de uma maneira ampla, foi dar tratamento eletrônico e lexicográfico aos dados do ALiB referentes às respostas dadas pelos informantes circunscritos na rede de pontos do interior da região Norte do Brasil, o banco de dados em *XML* também foi projetado para exibir informações em formato lexicográfico. Desse modo, a primeira etapa metodológica da pesquisa foi organizar os dados dialetais de forma lexicográfica em uma estrutura *XML*, conforme ilustrado na figura a seguir:

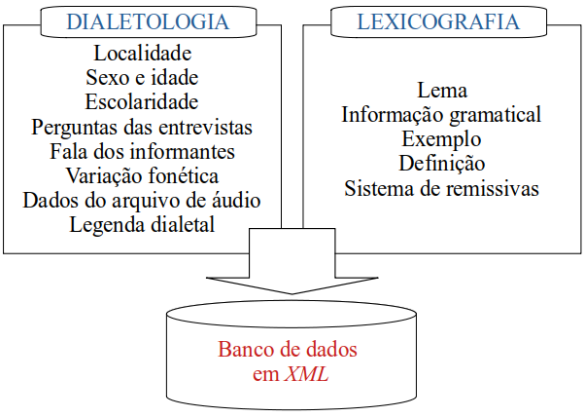
---

7 Para mais informações acesse o website em: https://alib.ufba.br/. Acesso em: 17 set. 2023.

8 Além do aspecto diatópico, predominante, a seleção dos informantes do Projeto ALiB também considerou a perspectiva diastrástica, dafásica e diageracional.

9 O armazenamento dos dados ocorreu a partir da transcrição do áudio das entrevistas nas *tags* do arquivo *XML*.

Figura 1: Primeira etapa metodológica da pesquisa



Fonte: Elaboração do autor.

Além dos tipos de informações que foram organizadas no arquivo *XML,* ilustradas na figura 1, foi preciso permitir a recuperação de informações levando em consideração que as perguntas das entrevistas do ALiB se repetem em cada nova localidade e a cada novo informante. Assim, uma identificação foi criada para cada resposta dos informantes com a finalidade de permitir a seleção de uma fala específica dentro do conjunto de todas as entrevistas. Essa identificação (entrada id) também possibilitou filtrar os resultados por uma das 14 áreas semânticas do Projeto ALiB, como é possível observar na figura a seguir:

Figura 2: Modelo de identificação das perguntas no banco de dados em *XML*

```
<entrada id="acid.geo.água.1">
<entrada id="fen.atm.chuva.12">
<entrada id="corpo.hum.membros.125">
<entrada id="jog.inf.grupo.12430">
```

Fonte: Elaboração do autor.

Constata-se, por meio da figura 2, que para cada entrada há uma identificação escrita entre aspas e formada por três grupos de dados, a saber: i) área semântica – destacada na cor vermelha em que as abreviações utilizadas referem-se às áreas semânticas *acidentes geográficos, fenômenos atmosféricos, corpo humano, jogos e diversões infantis*; ii) subárea semântica – sinalizada na cor preta em que a abreviação utilizada refere-se às perguntas do QSL da subárea *água, chuva, membros e grupo*; iii) identificação numérica – demarcada na cor azul em que uma sequência de números foi utilizada para que cada identificação seja única no arquivo *XML*. Assim, essa organização permite que o motor de busca recupere informações do *corpus* de pesquisa a partir desses três grupos de dados que, na prática, funcionam como etiquetas para a localização de dados específicos.

Além da identificação, outros campos foram adicionados na estrutura do *XML* de modo a permitir o armazenamento de informações pertinentes ao estudo como, por exemplo, uma área para o registro de observações gerais que o pesquisador julgue importante no decorrer do processo de transcrição dos dados, bem como outros campos destinados para a especificação dos arquivos de áudio, além de um espaço para adicionar os dados responsáveis pela exibição da legenda dialetal. Assim, após um período de ajustes necessários para atender aos objetivos da pesquisa, o arquivo *XML* assumiu a seguinte estrutura, conforme apresentada na figura a seguir:

Figura 3: Estrutura do banco de dados em *XML*

```
1  <?xml version="1.0" encoding="utf8" ?>
2  <!DOCTYPE dicio SYSTEM "corpus-micro.dtd">
3
4  <dicio>
5        <entrada id="fauna.faun.insetos.1064" abc="m">
6              <lema>mosquito</lema>
7              <perg campo="Fauna" ref="QSL-88/ALiB">Como se chama aquele inseto
pequeno, de perninhas compridas, que canta no ouvido das pessoas, de noite?</perg>
8              <ex>Carapanã, musquito. (É a mesma coisa carapanã i musquito?) É a mesma
coisa. (É o mesmo bichinho?) É o mesmo bichinho.</ex>
9              <obs></obs>
10             <fone>musquito</fone>
11             <aud src="fala-id-1064.mp3" type="mp3">Nome do arquivo = 1:05:15</aud>
12             <ver name="carapanã" ref="fauna.faun.insetos.1534"/>
13             <info sexo="Masculino" escolaridade="fundamental" idade="jovem" >Nome
do informante, 29 anos.</info>
14             <lg ponto="4" cidade="São Gabriel da Cachoeira" estado="AM"/>
15             <gram>Substantivo masculino</gram>
16             <def>Inseto pequeno que voa e pica, semelhante ao carapanã.</def>
17             <map src="Mapa-1064" type="jpg"/>
18       </entrada>
19 </dicio>
```

Fonte: Elaboração do autor.

A figura 3 representa um conjunto de dados armazenados em formato *XML* referentes à pergunta 88 do QSL/ALiB. Os dados são adicionados em *tags* que são os compartimentos específicos do aquivo responsáveis por armazenar dados. As *tags* são organizadas em um formato arbóreo e se caracterizam pelos símbolos < > que significa a abertura da *tag* e </ > que indica seu fechamento. Desse modo, é possível aludir o processo de armazenamento de dados no arquivo *XML* com o ato de abrir uma gaveta (< >), guardar um tipo de dado e fechar a gaveta (</ >). Para melhor compreensão da estrutura do banco de dados em *XML* apresenta-se, a seguir, uma sucinta descrição dos elementos exibidos na figura 3:

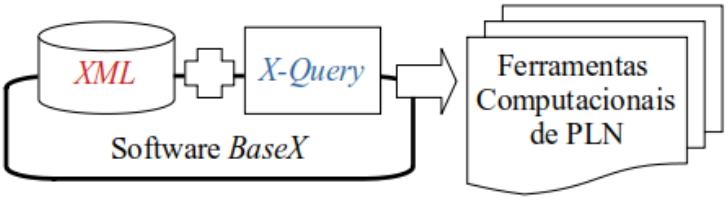
- Linha 1: Indica que o arquivo em questão se trata de um documento em formato *XML*, indicando sua versão (1.0) e a codificação dos caracteres presentes no documento (utf8);
- Linha 2: Informa que o arquivo foi construído em conjunto com um *Document Type Definition (DTD)*[10], que é utilizado para validar a estrutura do *XML*;
- Linha 4: *Tag* de abertura do banco de dados em *XML* que é considerada a raiz da estrutura do documento;
- Linha 5: *Tag* de abertura da *entrada* formada por um bloco de dados com uma identificação (fauna.faun.insetos.1064);
- Linha 6: *Tags* para armazenar o lema;
- Linha 7: *Tags* que exibem os dados da pergunta do QSL/ALiB;
- Linha 8: *Tags* para o exemplo lexicográfico, ou seja, a fala do informante;
- Linha 9: *Tags* reservadas para o registro de observações do pesquisador que possam ocorrer durante a transcrição das entrevistas;
- Linha 10: *Tags* para armazenar as variações fonéticas dos lemas;

[10] Arquivo com extensão *.dtd* em que são escritas as regras utilizadas para garantir a formatação correta do *XML*. O *DTD* escaneia o *XML* e caso encontre *tags* mal formatadas que possam resultar em erros durante o processo de recuperação de informações, um aviso é exibido ao usuário indicando a linha em que o erro se encontra.

- Linha 11: *Tags* destinadas para descriminação das informações do arquivo de áudio das entrevistas;
- Linha 12: *Tags* responsáveis pelo funcionamento do sistema de remissivas;
- Linha 13: *Tags* que especificam os dados do informante;
- Linha 14: *Tags* utilizadas para exibir as informações sobre o local da entrevista;
- Linha 15: *Tags* que têm a função de exibir a informação gramatical do lema;
- Linha 16: *Tags* usadas para a escrita da definição lexicográfica;
- Linha 17: *Tags* que guardam os dados da legenda dialetal;
- Linha 18: *Tag* de fechamento desse conjunto de dados. Corresponde à *tag* de abertura da linha 5;
- Linha 19: *Tag* de fechamento do banco de dados. Corresponde à *tag* de abertura da linha 4.

Após a estruturação do banco de dados em *XML* foi possível proceder com a etapa do armazenamento das informações nas respectivas *tags*. Esse trabalho levou muitos meses para ser concluído, tendo em vista a grande extensão do *corpus* das entrevistas do Projeto ALiB, referente à rede de pontos do interior da região Norte. Todavia, a partir de uma amostra de dados transcritos no arquivo *XML* foi possível desenvolver as primeiras ferramentas de PLN caracterizando, dessa forma, a segunda etapa metodológica da pesquisa, conforme ilustrado na figura a seguir:

Figura 4: Segunda etapa metodológica da pesquisa



Fonte: Elaboração do autor.

Observa-se, mediante a figura 4, que a segunda etapa metodológica tem como base o uso combinado do banco de dados em *XML* com as expressões *X-Query*, por meio do software *BaseX*[11] que, entre outras funções, permite a manipulação de dados a partir da escrita de linhas de código que funcionam como mecanismos de busca, selecionando e exibindo informações ao usuário de acordo com os comandos escritos na linguagem de consulta *X-Query*. Esse procedimento pode ser compreendido como uma etapa de desenvolvimento de ferramentas computacionais, ou seja, linhas de código que são escritas no editor de texto do *BaseX* para atender as necessidades de recuperação de informação de uma dada pesquisa científica. Desse modo, essas ferramentas computacionais desempenham um papel importante para o processamento de dados linguísticos, pois podem ser projetadas para realizarem tarefas de modo personalizado.

2.2. PROCESSAMENTO DOS DADOS

No intuito de ilustrar o funcionamento das ferramentas de PLN construídas para o processamento de dados dialetais e lexicográficos, três perguntas foram elaboradas cujas respostas

[11] Para mais informações acesse: https://basex.org/. Acesso em: 18 set. 2023.

podem ser encontradas no banco de dados em *XML* e, para cada pergunta, uma expressão *X-Query* foi escrita. As perguntas elaboradas foram:

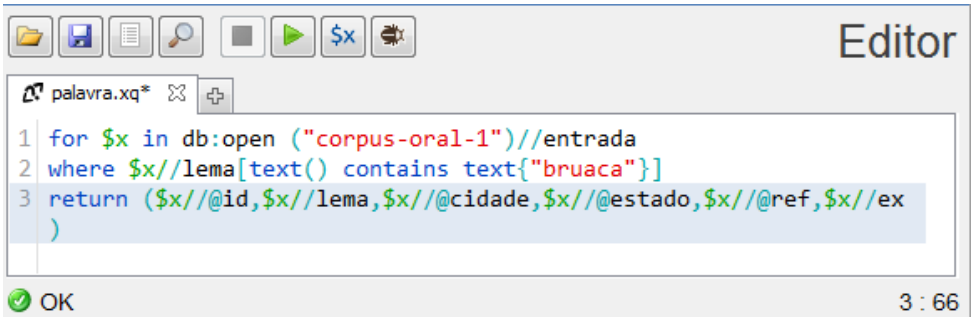
i) Houve menção da unidade lexical *bruaca* nas entrevistas realizadas nas localidades da rede de pontos do interior da região Norte do Brasil?

ii) Quais foram as respostas dadas pelos informantes referentes à área semântica *Atividades agropastoris*?

iii) Quais foram os resultados referentes à pergunta 88 – *Como se chama aquele inseto pequeno, de perninhas compridas, que canta no ouvido das pessoas, de noite?* – (Comitê Nacional..., 2001, p. 28) do QSL-ALiB, na cidade de Oiapoque/AP?

Desse modo, para responder à pergunta "Houve menção da unidade lexical *bruaca* nas entrevistas realizadas nas localidades da rede de pontos do interior da região Norte do Brasil?" foi preciso escrever, no editor do software *BaseX*, a seguinte *X-Query*:

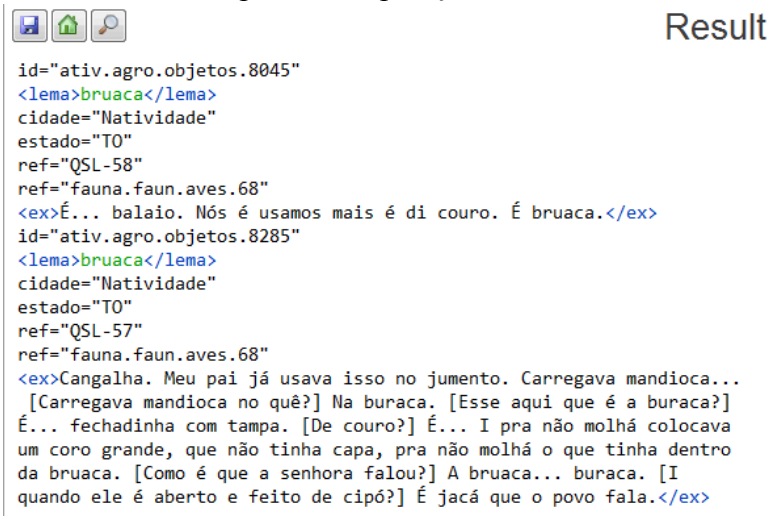
Figura 5: Expressão *X-Query* para recuperar a unidade lexical *bruaca*



Fonte: Software *BaseX*.

Na figura 5, é possível identificar três linhas de código que orientam as ações que a máquina deve executar. Para tanto, a primeira linha cria uma variável ($x) que armazenará, temporariamente, os dados que serão processados e exibidos para o usuário. Nessa linha há, ainda a instrução para se abrir o banco de dados nomeado como *corpus-oral-1* e acessar o conteúdo de todas as entradas. A segunda linha, por sua vez, orienta o sistema para recuperar o conteúdo textual *bruaca* em todos os *lemas* do banco de dados. Por fim, a terceira linha orienta quais *tags* deverão ser exibidas nos resultados que, nesse caso, são formadas pela identificação ($x//@id), lema ($x//lema), cidade ($x//@cidade), estado ($x//@estado), pergunta do QSL ($x//@ref) e exemplo ($x//ex). Os resultados da solicitação, discriminada na figura 5, podem ser visualizados na figura 6:

Figura 6: Resultados para a recuperação da unidade lexical *bruaca*

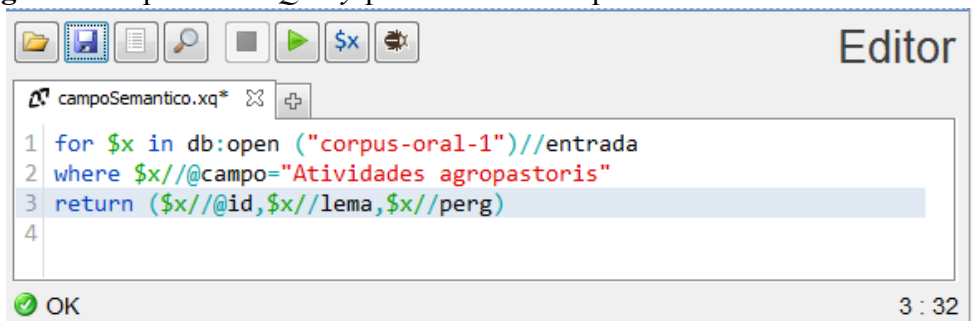


Fonte: Software *BaseX*.

As informações recuperadas do banco de dados e, exibidas na figura 6, indicam que o item lexical *bruaca* figura apenas duas vezes como *lema*, sendo mencionado como resposta para a pergunta 57 – *Como se chama aqueles objetos de vime, de taquara, de cipós trançado(s), para levar batatas (mandioca, macaxeira, aipim, etc.), no lombo do cavalo ou do burro?* – e 58 – *E quando se usam objetos de couro, com tampa, para levar farinha, no lombo do cavalo ou do burro? Mostrar gravura* – (Comitê Nacional..., 2001, p. 26), do QSL/ALiB. Além disso, ambas menções ocorreram na cidade de Natividade/TO e a fala completa dos informantes pode ser observada nas *tags* <ex> </ex>. Vale destacar que, caso o pesquisador deseje identificar que tipo de informante forneceu essas respostas, o comando adicional *and* deverá ser escrito ao final da linha dois da expressão *X-Query* apresentada na figura 5, seguido da variável ($x) e das respectivas *tags* que armazenam as informações relacionadas ao sexo (//@sexo) e idade (//@idade). Isso mostra como as expressões *X-Query* são versáteis para buscar e exibir os conteúdos das *tags* do *XML*.

Para responder à pergunta "Quais foram as respostas dadas pelos informantes referentes à área semântica *Atividades agropastoris*?" foi preciso escrever, no editor do *BaseX*, a seguinte expressão *X-Query*:

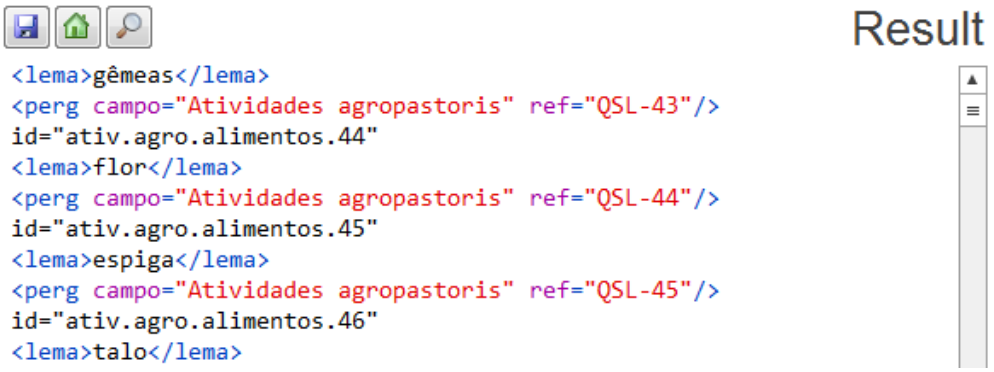
Figura 7: Expressão *X-Query* para filtrar dados por meio de uma área semântica



Fonte: Software *BaseX*.

É possível constatar, mediante a figura 7, que a primeira linha da *X-Query* não foi alterada, pois o banco de dados utilizado é o mesmo para todas as recuperações de informações. Desse modo, somente as demais linhas devem ser alteradas de acordo com o tipo de processo a ser realizado pela máquina. Para tanto, a segunda linha orienta o sistema para recuperar todos os dados referentes à área semântica *Atividades agropastoris*, enquanto a terceira linha, por sua vez, indica quais *tags* do banco de dados devem ser exibidas nos resultados, isto é, a identificação ($x//@id), o lema ($x//lema) e a pergunta do QSL/ALiB ($x//perg). Os resultados podem ser conferidos na figura a seguir:

Figura 8: Resultados para a recuperação de dados da área semântica *Atividades agropastoris*

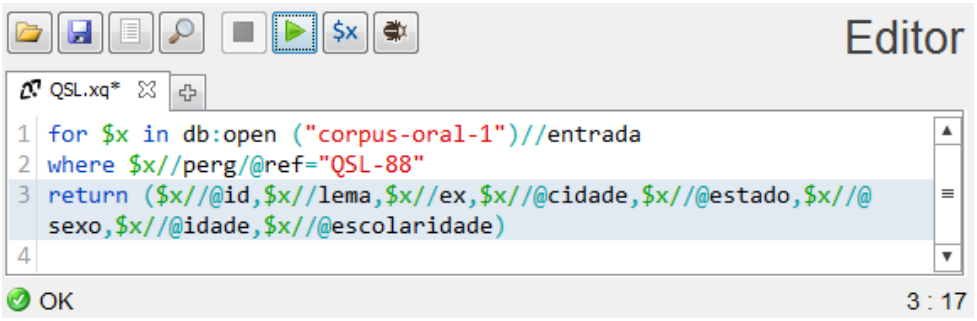


Fonte: Software *BaseX*.

Levando em consideração que os resultados exibidos pela solicitação apresentada na figura 7 são extensos, a figura 8 ilustra apenas três grupos de dados conforme a orientação dos comandos escritos na terceira linha da expressão *X-Query* (figura 7). Desse modo, é possível observar os lemas *gêmeas, flor* e *espiga* seguidos da indicação da área semântica em questão e do número da pergunta do QSL correspondente. Além disso, as *tags* que armazenam a identificação de cada conjunto de informações recuperadas exibem a área semântica (Atividades agropastoris), a subárea (alimentos) e a numeração (44, 45, 46) de cada conjunto de dados processados pela máquina.

Para responder à pergunta "Quais foram os resultados referentes à pergunta 88 – *Como se chama aquele inseto pequeno, de perninhas compridas, que canta no ouvido das pessoas, de noite?* – (Comitê Nacional..., 2001, p. 28) do QSL-ALiB, na cidade de Oiapoque/AP?" foi preciso escrever, no editor do *BaseX*, a seguinte expressão *X-Query*:

Figura 9: Expressão *X-Query* para selecionar dados referente à pergunta 88 do QSL/ALiB



Fonte: Software *BaseX*.

Novamente, é possível identificar, por meio da figura 9, que a primeira linha não foi alterada. Porém, a segunda linha foi editada e instrui o software para recuperar todas as informações armazenadas no banco de dados referente à pergunta 88 do QSL/ALiB. A terceira linha, ao seu turno, especifica quais *tags* devem ser exibidas ao usuário que, nesse caso, são compostas pela identificação ($x//@id), lema ($x//lema), exemplo ($x//ex), cidade ($x//@cidade), estado ($x//@estado), sexo ($x//@sexo), idade ($x//@idade) e escolaridade ($x//@escolaridade). Os resultados podem ser conferidos na figura a seguir:

Figura 10: Resultados para a recuperação de dados referente à pergunta 88 do QSL/ALiB

```
id="fauna.fauna.insetos.3965"
<lema>pernilongo</lema>
<ex>Pernilongo. Carapanã. [É igual?] É  mesma coisa. [Qual cê fala
mais?] Carapanã.</ex>
cidade="Humaitá"
estado="AM"
sexo="M"
idade="J"
escolaridade="F"
id="fauna.faun.insetos.4206"
<lema>carapanã</lema>
<ex>Carapanã ou então... como que é... mais provável é carapanã.</ex>
cidade="Humaitá"
estado="AM"
sexo="F"
idade="J"
escolaridade="F"
```

Fonte: Software *BaseX*.

Por se tratar de um conjunto de dados extenso, os resultados apresentados na figura 10 se limitam a dois grupos de informações, ou seja, o primeiro grupo representado pelo lema *pernilongo* e o segundo grupo pelo lema *carapanã*. É possível, ainda, observar a identificação numérica de cada lema, o contexto de fala do informante ao responder à pergunta 88 do QSL, bem como a especificação da cidade (Humaitá) e do estado (AM) em que as entrevistas foram realizadas, além da exibição dos dados do informante, isto é, o sexo (M e F), a idade (J)[12] e a escolaridade (F)[13].

3. ANÁLISE

O processamento de dados demonstrado no item anterior não teve como objetivo tecer discussões linguísticas relacionadas à Dialetologia e/ou à Lexicografia, mas ilustrar como foi possível realizar tarefas de PLN a partir de um *corpus* estruturado em formato *XML*. Nesse sentido, é importante destacar que a forma como as *tags* foram organizadas no banco de dados permitiu viabilizar a recuperação das informações que são de interesse da pesquisa. Outro fator importante que deve ser mencionado em relação ao processo de construção do banco de dados em *XML* é a liberdade que o pesquisador tem para organizar as informações hierarquicamente de acordo com os objetivos do projeto. Nesse sentido, as *tags* podem ser nomeadas da maneira que o usuário desejar, ou seja, não há um padrão rígido para a construção e nomeação desses campos e essa flexibilidade permite uma organização mais intuitiva, aos olhos humanos, de toda a estrutura do banco de dados.

É preciso destacar, por sua vez, que as expressões *X-Query* são capazes de realizar diversificadas buscas no *XML,* atendendo as necessidades relacionadas à manipulação eletrônica de dados no âmbito da Dialetologia e da Lexicografia. No entanto, para cada tipo de tarefa de PLN uma diferente *X-Query* deve ser escrita e esse procedimento, no início, pode ser um desafio para o pesquisador que está desbravando esse conteúdo em particular.

Há de se frisar, ainda no campo dos desafios, que compreender como funciona o software *BaseX* pode ser um obstáculo inicial. Porém, com o auxílio de uma amostra de dados em *XML* para testes que acompanha o software e, munido da documentação disponível no website do desenvolvedor, as dificuldades tendem a diminuir em algumas semanas de treinamento e o pesquisador poderá utilizar a plataforma para gerenciar seu próprio arquivo *XML*.

Em síntese, o uso de metodologias computacionais destinadas ao desenvolvimento das próprias ferramentas, a fim de atender aos objetivos de qualquer projeto científico, perpassa por questões que envolvem a disposição e o tempo do pesquisador para aprender os conteúdos pontuais do universo da Ciência da Computação que serão utilizados na pesquisa, bem como exige a consultoria de um especialista para nortear os caminhos que deverão ser percorridos durante essa jornada.

O resultado de todo esse empreendimento interdisciplinar é visto nas diversificadas possibilidades de processamentos de dados em um determinado projeto, além de possibilitar a construção de produtos linguísticos eletrônicos que podem acessar variados bancos de dados em *XML*, tendo em vista que esse formato é compatível com outros sistemas computacionais. Destaca-se, ainda, que o uso integrado de bancos de dados é, atualmente, um dos métodos que a Lexicografia Eletrônica tem explorado para exibir informações tratadas lexicograficamente ao consulente por meio de plataformas on-line.

---

[12] A abreviação *J* refere-se ao informante jovem (entre 15 e 30 anos). Os informantes idosos (entre 50 e 65 anos) foram demarcados no banco de dados com a abreviação *I*.

[13] A abreviação *F* refere-se ao ensino fundamental de escolaridade e é um dos critérios para a escolha dos informantes nos municípios do interior. Porém, nas capitais, além de um grupo de informantes com ensino fundamental, também há outro grupo de entrevistados que possuem o nível superior de escolaridade.

CONCLUSÃO

Investir no uso de metodologias computacionais como, por exemplo, *XML* e *X-Query* permite a construção de ferramentas computacionais próprias e modeladas para atender aos objetivos de uma dada pesquisa científica, viabilizando a manipulação eletrônica de dados, bem como a integração do *corpus* de estudo com outros projetos, tendo em vista que os bancos de dados em *XML* podem integrar outros sistemas informatizados.

Nesse sentido, é importante que acadêmicos e professores/pesquisadores da área de Letras tenham acesso às metodologias computacionais de análises lexicais, sobretudo às técnicas destinadas ao desenvolvimento das próprias ferramentas computacionais, permitindo uma ampliação dos horizontes de pesquisa e produção de produtos digitais inovadores. Em síntese, lançar mão dessas técnicas propicia o desenvolvimento de estudos linguísticos sob uma perspectiva estatística, bem como auxilia na observação de regularidades, no teste de hipóteses e na recuperação eletrônica de informação que pode ser útil em qualquer área do conhecimento.

REFERÊNCIAS

BERBER-SARDINHA, Tony. *Linguística de Corpus.* Barueri: Manole, 2004.

BLOOMFIELD, Leonard. *Language*. New York: Henry Holt, 1933.

CHAMBERS, Jack; TRUDGILL, Peter. *La dialectología.* Madrid: Visor Libros, 1994.

CHOMSKY, Noam. *Aspects of the theory of syntax.* Cambridge-MA: MIT Press, 1965.

COMITÊ NACIONAL DO PROJETO ALIB. *Atlas Lingüístico do Brasil:* questionário 2001. Londrina: EDUEL, 2001.

O'KEEFFE, Anne; MCCARTHY, Michael. What are corpora and how have they evolved? *In:* O'KEEFFE, Anne; MCCARTHY, Michael (ed.). *The Routledge handbook of corpus linguistics.* London; New York: Routledge, 2010. p. 3-10.

ORLANDI, Eni Puccinelli. *O que é linguística.* 2. ed. São Paulo: Brasiliense, 2009.

SANTOS JUNIOR, Jorge Luiz Nunes dos. *Interfaces entre Lexicografia e Dialetologia:* por um protótipo de vocabulário dialetal eletrônico da região Norte do Brasil, 2023. Tese (Doutorado em Letras) – Universidade Federal de Mato Grosso do Sul, Três Lagoas, 2023.

SRINIVASA-DESIKAN, Bhargav. *Natural Language Processing and Computational Linguistics:* A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Birmingham: Packt, 2018.

TARP, Sven. Connecting the dots: tradition and disruption in Lexicography. *Lexikos*, v. 29, p. 224-249, 2019. Disponível em: http://lexikos.journals.ac.za. Acesso em: 17 set. 2023.

WALMSLEY, Priscilla. *XQuery*: Search Across a Variety of XML Data. 2. ed. Sebastopol-CA: O'Reilly, 2015.