Desenvolvimento de ferramentas computacionais para o processamento de dados dialetais e lexicográficos

Computational tools development for the dialectal and lexicographical data processing

Jorge Luiz Nunes dos Santos Junior *[1]

[1] *Universidade Federal de Mato Grosso do Sul, Programa de Pós-Graduação em Letras, Três Lagoas, MS, Brasil.*

Resumo

Este trabalho situa-se na intersecção da Linguística de Corpus (O'KEEFFE; MCCARTHY, 2010); Linguística Computacional (KEDIA; RASU, 2020; SRINIVASA-DESIKAN, 2018; MANNING, 2008; MANNING; SCHUTZE, 1999; CHOMSKY, 1965); Dialetologia (CARDOSO, 2010; RADTKE; THUN, 1996; CHAMBERS; TRUDGILL, 1994) e Lexicografia (TARP, 2008, 2011, 2015; FUERTES-OLIVEIRA; BERGENHOLTZ, 2015; LEROYER, 2011). Tem-se como objetivo apresentar o desenvolvimento de ferramentas computacionais capazes de processar dados de natureza dialetal e lexicográfica a partir de uma metodologia que dispensa a contratação de serviços de programação, convidando o pesquisador a estudar os recursos informáticos necessários para realizar a manipulação automática de informações em um banco de dados. Para tanto, o *corpus* utilizado foi o do Projeto Atlas Linguístico do Brasil (COMITÊ NACIONAL DO PROJETO ALIB, 2001), relativo aos municípios do interior, da rede de pontos do ALiB, na região Norte do país. A construção desses pequenos programas foi motivada, principalmente, por duas razões: i) fornecer tratamento lexicográfico e eletrônico aos dados dialetais do ALiB; ii) desenvolver as próprias ferramentas computacionais para atender aos objetivos da pesquisa de Doutorado em andamento à qual este artigo se vincula. Desse modo, um banco de dados em *Extensible Markup Language* (XML) foi construído para armazenar as informações dialetais em formato lexicográfico e, a partir da execução de linhas de código, foi possível recuperar dados específicos do corpus de maneira eletrônica, além de filtrar os resultados a partir das variantes 'sexo', 'idade' e 'localidade', presentes nos dados do corpus do ALiB.

Palavras-chave: Dialetologia. Lexicografia. Ferramentas computacionais. Linguagens de programação. Banco de dados.

Abstract

This paper is situated at the intersection of Corpus Linguistics (O'KEEFFE; MCCARTHY, 2010); Computational Linguistics (KEDIA; RASU, 2020; SRINIVASA-DESIKAN, 2018; MANNING, 2008; MANNING; SCHUTZE, 1999; CHOMSKY, 1965); Dialectology (CARDOSO, 2010; RADTKE; THUN, 1996; CHAMBERS; TRUDGILL, 1994) and Lexicography (TARP, 2008, 2011, 2015; FUERTES-OLIVEIRA; BERGENHOLTZ, 2015; LEROYER, 2011). It aims to present the development of computational tools capable of processing dialectal and lexicographic data using a methodology that does not require the hiring of programming services, inviting the researcher to study the necessary computer resources to perform an automatic manipulation of information in a database. For this purpose, the *corpus* used was *Atlas Linguístico do Brazil Project* (COMITÊ NACIONAL DO PROJETO ALIB, 2001) relating to the interior municipalities from the ALiB, network, pointed out in the country's North region. The construction of these small programs was mainly motivated by two reasons: i) provide lexicographical and electronic treatment to ALiB dialect data; ii) develop their own computational tools to meet the Doctoral research goals in progress, to which this article is linked. Thus, a database in Extensible Markup Language (XML) was built to store dialectal information in lexicographical format, and through the execution of code lines, it was possible to electronically retrieve specific data from the *corpus* and filter the results based on 'gender', 'age', and 'location' variants present in the data from the ALiB *corpus*.

Keywords: Dialectology. Lexicography. Computational tools. Programming languages. Database.

Textolivre
Linguagem e Tecnologia

DOI: 10.1590/1983-3652.2023.42302

Seção:
Artigos

Autor Correspondente:
Jorge Luiz Nunes dos Santos Junior

Editor de seção:
Danierveli Pereira
Editor de layout:
Thaís Coutinho

Recebido em:
30 de dezembro de 2022
Aceito em:
8 de março de 2023
Publicado em:
11 de abril de 2023

*Email: jorgesantosjunior@gmail.com

1 Introdução

Com o desenvolvimento cada vez mais frenético das Ciências da Computação, programas de computador são criados para realizar diversas tarefas em todas as áreas do conhecimento. O crescimento da oferta de *softwares* nas variadas áreas do entretenimento, trabalho e pesquisa é resultado de uma evolução tecnológica que a cada dia lança uma novidade ou uma versão atualizada de produtos computacionais.

A partir desse crescimento tecnológico, atualmente é possível desenvolver programas para uma grande gama de atividades e isso tem revolucionado o fazer científico. No entanto, rapidez no processamento de dados não é o grande trunfo dessas máquinas, pois isso é o mínimo que um computador pode realizar. O maior ganho que se tem com a utilização de programas de computador, no âmbito da pesquisa científica, é a possibilidade de abertura de novos horizontes ao pesquisador. Trata-se da possibilidade de oferecer ângulos variados de observação e maneiras múltiplas de manipulação e transformação de dados.

Nesse sentido, é importante que o pesquisador saiba utilizar, ao menos em nível elementar, *softwares* que realizem o tratamento e o processamento de informações em sua área de atuação.

No âmbito da Linguística, por exemplo, diversas áreas utilizam dados compilados em um corpus a fim de atender os objetivos de determinado estudo como acontece, frequentemente, no campo da Lexicologia, Fraseologia, Terminologia, Lexicografia, Dialetologia, entre outras áreas. Nesse sentido, o estudioso interessado em realizar análises lexicais pode contar com programas de computador que executam algumas tarefas árduas do ponto de vista humano como é o caso do *AntConc*[1], *LancsBox*[2], *WordSmith Tools*[3] e *Sketch Engine*[4] que são programas que oferecem um pacote de ferramentas capazes de: i) organizar listas de palavras em ordem de frequência; ii) mostrar o contexto de uma palavra ou de uma expressão; iii) comparar as palavras de um corpus de estudo com dados de um corpus de referência; iv) realizar cálculos estatísticos sobre a disposição lexical em um corpus.

Outro programa muito utilizado em pesquisas lexicográficas é o *FieldWorks Language Explore*[5] *(FLEx)* que oferece recursos para a edição de verbetes a partir de dados inseridos no software, mediante critérios estabelecidos pela ferramenta. De acordo com os desenvolvedores, o *FLEx* permite, ainda, exportar os dados para uma plataforma externa, a fim de que o produto lexicográfico possa ser publicado em um *website*.

Todavia, com todas as funcionalidades que podem ser exploradas pelo pesquisador envolvido com temáticas linguísticas, o uso desses e/ou de outros *softwares* são limitados, ou seja, só é possível realizar tarefas para o qual o programa foi projetado.

Não há nada de errado nisso e é compreensível que um programa de computador funcione para realizar um limitado número de tarefas. Compreende-se, ainda, que as demandas por *softwares* surgem de acordo com os objetivos traçados no âmbito de cada pesquisa, sendo difícil prever todas as possibilidades de manipulação de dados em uma determinada área do conhecimento.

Nesse sentido, ocorrem com frequência situações em que muitas atividades que poderiam ser executadas por meio do computador acabam sendo realizadas manualmente, porque o *software* utilizado não atende a determinadas necessidades.

Nesses casos, uma possível solução para resolução do impasse seria o desenvolvimento de ferramentas computacionais[6] personalizadas, isto é, pequenos programas que executem as tarefas que o pesquisador necessita. Para tanto, a partir dos objetivos de uma pesquisa científica, o estudioso deve, em primeiro lugar, delimitar as ações que deverão ser processadas eletronicamente e, em se-

[1] *software* gratuito desenvolvido por Laurence Anthony na Universidade de Waseda, Japão. Disponível em: http://www.laurenceanthony.net/software/antconc/. Acesso em 10 out. 2022.

[2] *software* gratuito desenvolvido por Vaclav Brezina, William Platt e Tony McEnery na Universidade de Lancaster, Reino Unido. Disponível em: http://corpora.lancs.ac.uk/lancsbox/. Acesso em 10 out. 2022.

[3] *software* pago criado por Mike Scott e publicado pela *Lexical Analysis software e Oxford Universit Press*. Disponível em: https://lexically.net/wordsmith/. Acesso em 10/10/2022.

[4] *software* pago criado por Adam Kilgarriff; Pavel Rychlý e desenvolvido pela *Lexical Computing CZ s.r.o.* Disponível em: https://www.sketchengine.eu/. Acesso em 10 out. 2022.

[5] *software* gratuito desenvolvido pela *SIL International*. Disponível em: https://software.sil.org/fieldworks/. Acesso em 10 out. 2022.

[6] O termo *ferramentas computacionais* é usado neste artigo como referência ao uso de linguagens de programação para a execução de tarefas específicas em um ambiente informatizado.

gundo lugar, investir em conhecimento técnico que o capacitará para a construção dessas soluções informáticas.

Não há dúvidas que adquirir o conhecimento mínimo do universo das linhas de código[7] representa um grande desafio a ser vencido, sobretudo ao estudioso desprovido de conhecimentos sobre programação. Todavia, antes de recorrer a um profissional do ramo da Informática para terceirizar esse laborioso serviço é possível, a partir da metodologia apresentada e discutida neste artigo, que o linguista reconsidere essa terceirização e se anime para enveredar por caminhos no universo das linguagens artificiais que foram criadas pelo homem, a fim de que o computador possa executar instruções.

Desse modo, este artigo pretende compartilhar a experiência que está sendo vivenciada no âmbito de uma pesquisa de Doutorado que tem como objetivo mais amplo elaborar o protótipo do Vocabulário Dialetal da região Norte do Brasil (VoDiNorte) a partir de dados do Projeto Atlas Linguístico do Brasil (COMITÊ NACIONAL DO PROJETO ALIB, 2001), referente aos municípios do interior da rede de pontos do ALiB na região Norte do Brasil. Para tanto, esses dados estão recebendo um tratamento lexicográfico e eletrônico. Neste sentido, ferramentas computacionais foram desenvolvidas especialmente para atender as necessidades de manipulação e transformação das informações dialetais que estão sendo armazenadas em um banco de dados em *XML*[8]. Em síntese, o objetivo deste trabalho é apresentar o desenvolvimento de ferramentas computacionais capazes de processar dados dialetais e lexicográficos, além de discutir os benefícios que o pesquisador tem ao optar pela criação de suas próprias soluções informatizadas em uma pesquisa cientifica.

2 Referencial teórico

Este artigo fundamenta-se em disciplinas que se relacionam de tal modo que fica claro o caráter interdisciplinar das áreas do conhecimento aqui mobilizadas. Desse modo, o estudo conta com o aporte da Linguística de *Corpus*, Linguística Computacional, Dialetologia e Lexicografia.

2.1 Linguística de *Corpus*

É possível identificar, em meados do século XVIII, estudos de caráter linguístico que utilizavam métodos para indexar listas de palavras a partir de um conjunto textual. Isso significa que a metodologia conhecida atualmente como Linguística de Corpus surgiu, em tempos remotos, da necessidade que os estudiosos da Bíblia Sagrada tinham para elaborar uma relação das palavras contidas na Bíblia, juntamente com a indicação de onde elas ocorriam. Essa prática manual caracteriza-se pela construção de concordâncias, ou seja, a organização de um texto com indicativos de frequência e a possibilidade de observar cada ocorrência juntamente com o seu contexto. Esse árduo trabalho contava com equipes que chegavam a centenas de homens com a finalidade de criar as denominadas *listas de palavras* que, nos dias atuais, podem ser geradas por qualquer pessoa que saiba utilizar, elementarmente, um Concordanceador[9] (O'KEEFFE; MCCARTHY, 2010, p. 3).

Observa-se, desse modo, que a metodologia da Linguística de Corpus é anterior à era dos computadores eletrônicos, que teve seu início marcado por volta dos anos 1940 e que se encontra, atualmente, em um crescimento exponencial, sem precedentes e de difícil acompanhamento por grande parcela dos seres humanos. Esse crescimento pode ser constatado, como mencionado anteriormente, na variedade de programas computacionais que são lançados e/ou atualizados constantemente e em um ritmo acelerado.

Como todas as áreas do conhecimento, a Linguística se beneficiou do desenvolvimento informático, o que pode ser observado na Linguística de *Corpus* que, aliada ao desenvolvimento de *softwares* capazes de processar as línguas naturais, ampliou significativamente seu leque de atuação. Desse

[7] Instruções fornecidas ao computador em uma linguagem que a máquina possa entender. As linhas de código estão presentes em várias linguagens de programação, incluindo as expressões *X-Query* que foram utilizadas na construção das ferramentas computacionais apresentadas neste artigo.

[8] Acrônimo do termo em inglês Extensible Markup Language. É uma linguagem de marcação capaz estruturar dados em um formato arbóreo, permitindo a recuperação de informações eletronicament

[9] Ferramenta encontrada nos pacotes de programas como AntConc capaz de indexar as palavras de um corpus em ordem de frequência, além de possibilitar a observação de seu contexto.

modo, a Linguística de *Corpus* não é utilizada apenas em estudos de ordens lexicais e gramaticais, mas também tem servido de referencial metodológico para pesquisas empíricas em áreas como o ensino e aprendizagem de línguas, análise do discurso, estilística, linguística forense, pragmática e sociolinguística (O'KEEFFE; MCCARTHY, 2010, p. 7). Uma aplicação desse campo metodológico no ramo da tradução pode ser observada, por exemplo, na ferramenta de tradução automática *Linguee*[10] que mostra ao usuário alguns contextos de uso da palavra pesquisada nos dois idiomas, isto é, na língua de partida e de chegada.

Observa-se, desse modo, que o desenvolvimento computacional tem contribuído para ampliar as possibilidades de investigações na área da Linguística, especialmente quando utilizadas para aperfeiçoar os métodos da Linguística de *Corpus*. Esse desenvolvimento colaborou para a constituição de uma área, interdisciplinar, que abrange uma gama de investigações no âmbito da linguagem denominada Linguística Computacional.

2.2 Linguística Computacional

A Linguística Computacional se apresenta como um campo de estudo da linguagem que busca analisar seu objeto de estudo por meio do auxílio de programas computacionais. Nesse sentido, estabelece uma interface com o Processamento (Automático) da Linguagem Natural (PLN) que, por sua vez, consiste em uma área interdisciplinar – ligada às Ciências da Computação e à Inteligência Artificial – que busca desenvolver ferramentas capazes de manipular a linguagem natural, tendo em vista que o computador entende uma linguagem diferente da humana.

Desse modo, para que um software consiga processar dados linguísticos é preciso dizer à máquina o que ela deve fazer. Em síntese, o PLN – acrônico do termo inglês *Natural Language Processing (NLP)* – configura-se, de acordo com Kedia e Rasu (2020, p. 7, tradução nossa), numa "[...] interdisciplinary area of research aimed at making machines understand and process human languages"[11]. Os autores ainda pontuam que o PLN é uma área que está evoluindo rapidamente e sua aplicação pode ser vista em programas desenvolvidos para a indústria computacional como, por exemplo, Alexia, Google Tradutor e *chatbots*.

Atualmente, a temática que envolve a inteligência artificial está em alta, desde que a empresa *OpenAI*[12] disponibilizou ao público o *ChatGPT*. Sem entrar em questões de ordem técnica é possível afirmar que um dos recursos mais impactantes nesse tipo de software é a capacidade de compreender a linguagem natural podendo estabelecer um diálogo com o usuário em diferentes idiomas. Essa amplitude linguística é resultado das pesquisas em PLN que buscam "compreender" como as línguas naturais funcionam.

É importante destacar que entender o funcionamento de uma língua natural é demasiadamente complexo. Um falante nativo é capaz de perceber as nuances sutis em uma língua como, por exemplo, questões relacionadas à pragmática ou à ambiguidade. Todavia, esses usos linguísticos desafiam os *chatbots* que recorrem a modelos de língua baseados em dados probabilísticos para formar sentenças cada vez mais próximas da linguagem humana.

É nesse sentido que as contribuições da gramática gerativa de Chomsky (1965, p. 15) auxiliam na compreensão da sintaxe das línguas naturais. O estudo desse linguista resultou em um sistema de regras, que descrevem as estruturas sintagmáticas em uma frase, que é o ponto de partida para uma sistematização probabilística do léxico de um idioma por meio de algoritmos de computador. Isso significa que para o desenvolvimento de ferramentas de PLN uma abordagem estatística é adotada para identificar padrões que ocorrem no uso da linguagem (MANNING; SCHUTZE, 1999, p. 4).

Além disso, dentre os métodos de PNL é possível destacar a organização de dados de forma estruturada como, por exemplo, em *XML*, com a finalidade de facilitar a recuperação de uma informação que atenda a uma necessidade pontual de um usuário em um momento específico (MANNING, 2008, p. 184). Assim, o ato de instruir o computador, por meio de uma linguagem de marcação, para que a

[10] A ferramenta pode ser acessada em https://www.linguee.com.br/. Acesso em: 10 out. 2022.
[11] "[...] área interdisciplinar de pesquisa que visa a fazer com que as máquinas entendam e processem as linguagens humanas" (KEDIA; RASU, 2020, p. 7).
[12] Para mais informações acesse https://openai.com/. Acesso em 16 mar. 2023.

máquina possa recuperar um conteúdo armazenado em uma *tag XML* consiste em uma metodologia de PLN usada por necessidades de pesquisas diversas e oriundas de qualquer área do conhecimento. Desse modo, observa-se que, mesmo tendo objetivos distintos, a Linguística Computacional e o PLN são áreas que se complementam em uma relação de interconexão.

Em suma, Srinivasa-Desikan (2018, p. 11, tradução nossa) resume que a Linguística Computacional "[...] is the study of linguistics from a computational perspective. This means using computers and algorithms to perform linguistics tasks such as marking your text as a part of speech (such as noun or verb), instead of performing this task manually"[13]. Além dessa síntese o exemplo apresentado por esse autor ilustra uma das maneiras de automatizar tarefas normalmente realizadas pelo homem. Destaca-se, todavia, que o escopo da Linguística Computacional vai além da marcação morfossintática[14] de um *corpus* de modo que as pesquisas linguísticas, sob a perspectiva computacional, podem ser desenvolvidas, praticamente, em qualquer área que envolva a linguagem natural como, por exemplo, no campo da Lexicografia e da Dialetologia.

2.3 Dialetologia

Segundo Cardoso (2010, p. 15), a Dialetologia é um "[...] um ramo dos estudos linguísticos que tem por tarefa identificar, descrever e situar os diferentes usos em que uma língua se diversifica, conforme a sua distribuição espacial, sociocultural e cronológica". Nesse sentido, os estudos dialetais desempenham uma importante tarefa na descrição da norma lexical em uma determinada região geográfica a partir de critérios bem definidos do ponto de vista geolinguístico.

Assim, a coleta de dados *in loco* se torna uma realidade no desenvolvimento de pesquisas dialetais, feitas a partir de entrevistas com um determinado grupo de informantes que são escolhidos segundo alguns critérios que, posteriormente, subsidiam a comparabilidade dos dados. Dessa forma, as denominadas variáveis dialetais delimitam o *corpus* que será compilado, permitindo observar dados sob ângulos diversos, o que constitui um fator importante para se estudar a norma lexical de uma dada região.

Na fase da denominada Dialetologia tradicional, considerava-se que o informante ideal deveria ser do sexo masculino, idoso, de baixa escolaridade e sedentário (CHAMBERS; TRUDGILL, 1994, p. 84), ou seja, um período em que as investigações eram de natureza monodimensionais, pois não consideravam possíveis variáveis sociais na delimitação do tipo de informante que seria selecionado para as entrevistas.

Atualmente, a coleta de dados busca reunir informações dialetais considerando um conjunto de variáveis como, por exemplo, relacionadas ao informante – masculino, feminino, idoso ou jovem – que diferenciam o grau de instrução – nível fundamental, médio ou superior – e a variável que estabelece o espaço geográfico a ser estudado – rural ou urbano. Acrescentam-se, ainda, aspectos sociais, culturais e de estilística que influenciam o uso da língua em situações de comunicação oral. Por considerar um número maior de variáveis, tais estudos representam a nova geração de investigações dialetais, denominada pluridimensional (RADTKE; THUN, 1996, p. 48-49) e contatual.

Nesse contexto, partindo do pressuposto metodológico de que a Dialetologia trabalha com dados organizados a partir de variáveis espaciais/geográficas, é importante que a manipulação dessas informações ocorra de modo automatizado, ou seja, por meio de ferramentas computacionais capazes de recuperar informações específicas em um banco de dados, como será detalhado neste artigo.

Acrescenta-se, ainda, que, além de automatizar a manipulação de dados referentes às variáveis estabelecidas em um estudo dialetal, o desenvolvimento de ferramentas computacionais pode atender a necessidades de outras áreas como, por exemplo, aquelas relacionadas ao labor lexicográfico.

[13] "[...] é o estudo da linguística a partir de uma perspectiva computacional. Isso significa usar computadores e algoritmos para realizar tarefas linguísticas, como marcar seu texto como parte da fala (como substantivo ou verbo), em vez de realizar essa tarefa manualmente" (SRINIVASA-DESIKAN, 2018, p. 11).

[14] Para maiores informações sobre marcação morfossintática acesse: https://www.cis.lmu.de/~schmid/tools/TreeTagger/. Acesso em: 10 out. 2022.

2.4 Lexicografia

A elaboração de uma obra lexicográfica exige uma metodologia rigorosa, a fim de garantir a qualidade ao produto final. Nesse sentido, é inviável produzir um dicionário, atualmente, sem o uso do computador. Nesta questão, em particular, é possível destacar as funcionalidades de programas que podem realizar determinadas tarefas no campo da Lexicografia, como é o caso dos *softwares* de análise lexical mencionados anteriormente. Além do mais, há de se considerar a possibilidade do desenvolvimento de ferramentas computacionais próprias para atender a objetivos específicos de projetos lexicográficos.

Essa necessidade parte do fato de que os *softwares* disponíveis no mercado, ou seja, aqueles que possuem uma interface para o manuseio do usuário com botões e campos para a inserção de dados, realizam apenas as tarefas para a qual foram projetados e, dessa forma, nem sempre irão atender plenamente a demanda dos pesquisadores e/ou lexicógrafos. Em contrapartida, saber programar tarefas úteis à Lexicografia, fazendo o uso de linguagens de programação, amplia as possibilidades de manipulação dos dados, tendo em vista que é o pesquisador o responsável por delimitar e editar, sempre que julgar necessário, as linhas de código responsáveis pela automação de tarefas.

Outro ponto que precisa ser destacado é o tipo de uso que a Lexicografia tem feito da tecnologia computacional nos últimos anos. Essa questão é importante, pois o futuro da Lexicografia é a produção de dicionários capazes de oferecer ao usuário a consulta a dados dinâmicos[15], ou seja, informações lexicográficas que são recuperadas, automaticamente, por meio de várias fontes e selecionadas segundo o perfil do usuário (TARP, 2011, p. 63). Isso significa que os dicionários poderão acessar um ou vários bancos de dados armazenados em seu próprio servidor e também terão a possibilidade de vasculhar a Internet em busca de informações que possam ser aproveitadas para enriquecer o verbete que está sendo exibido ao consulente.

No entanto, é preciso destacar que há uma diferença fundamental entre um dicionário publicado em uma plataforma eletrônica e uma obra lexicográfica que é projetada e desenvolvida para funcionar em plataformas on-line. Assim, os dicionários publicados em suporte eletrônico são concebidos a partir de metodologias provenientes da Lexicografia impressa, ao passo que os dicionários construídos somente a partir das modernas tecnologias computacionais são baseados nos preceitos da e-Lexicografia, que tem se dedicado a construir ferramentas de informação lexicográfica[16] que atendam as necessidades específicas de seus usuários.

Destaca-se, ainda, que o processo de concepção de dicionários on-line ancora-se em três elementos fundamentais e que se relacionam entre si, a saber: i) o usuário; ii) o acesso; iii) e os dados (LEROYER, 2011, p. 128). Nesse contexto, deve-se levar em conta a figura do usuário como um sujeito que possui um perfil, além de demandas variadas de pesquisa. Isso é importante para que se possa garantir a satisfação do usuário em sua consulta e, partindo desse princípio, projetar ferramentas de informação lexicográfica capazes de administrar os dados da triangulação citada por Leroyer (2011, p. 128), que poderá acessar variados bancos de dados (que podem estar espalhados pelo mundo) para exibir um verbete adequado a um determinado tipo de usuário.

Com base nesses pilares, Leroyer (2011, p. 129, tradução nossa) assim define Lexicografia:

> Lexicography is an integrated part of the social and information science paradigm and refers to the interdisciplinary discipline concerned with the study, design and development of functional tools aimed solely at the gratification of human information needs and problems[17].

[15] Os dados dinâmicos são aqueles que se adaptam ao tipo de usuário a partir de mecanismos que se destinam a identificar as necessidades de pesquisa do consulente, ou seja, são dados lexicográficos exibidos de acordo com o tipo de usuário. Em contrapartida, os dados estáticos se caracterizam por informações da microestrutura que são exibidas ao consulente em dicionários on-line e são sempre os mesmos. Esse é o modelo que nasceu na lexicografia impressa e tem sido usado em dicionários on-line de todo o mundo.

[16] Para Fuertes-Oliveira e Bergenholtz (2011, p. 2) e Tarp (2011, p. 69) as mudanças no campo da Lexicografia têm aberto uma discussão sobre o fato do termo dicionário não condizer, do ponto de vista semântico, com as obras que estão sendo desenvolvidas no âmbito da e-Lexicografia. Desse modo, esses autores sugerem a utilização de uma nomenclatura mais apropriada como, por exemplo, *ferramenta de informação lexicográfica, ferramenta de consulta lexicográfica ou e-ferramenta lexicográfica*.

[17] "A lexicografia é parte integrante do paradigma da ciência social e da informação e refere-se à disciplina interdisciplinar preocupada com o estudo, projeto e desenvolvimento de ferramentas funcionais voltadas exclusivamente para a satisfação das necessidades e problemas de informação humana" (LEROYER, 2011, p. 129)

Vale destacar que a natureza da Lexicografia é a interdisciplinaridade. Assim, o fato de aproximá-la como parte integrante da ciência social e da informação não anula as bases construídas ao longo da história de caráter linguístico. O que ocorre, atualmente, é a necessidade de complementar o escopo teórico e metodológico da Lexicografia, a fim de que as publicações de obras lexicográficas digitais tenham qualidade e não sejam uma simples compilação de dados feita de qualquer modo e que buscam apenas lucrar na Internet.

Para tanto, a Teoria Funcional da Lexicografia (TARP, 2008, p. 43) foi desenvolvida com o objetivo de dar sustentação teórica e metodológica às ferramentas de informação lexicográficas que têm a missão de suprir as necessidades de seus usuários. O termo funcional ou funções alude, dessa forma, ao serviço que todo bom dicionário deve prestar ao usuário:

> [...] una función lexicográfica puede definirse como la asistencia que presta una obra lexicográfica para satisfacer los tipos específicos de necesidades de información puntual que pueda tener un tipo específico de posible usuario en un tipo específico de situación extra-lexicográfica. La asistencia a que se refiere se logra por medio de los datos lexicográficos detenidamente preparados y hechos accesibles para su consulta[18] (TARP, 2015, p. 26, tradução nossa).

Observa-se, dessa maneira, que a Teoria Funcional da Lexicografia tem fundamentado a construção de dicionários on-line que têm utilizado os recursos informáticos para elaborar soluções lexicográficas inovadoras como, por exemplo, o projeto dos *Dicionários on-line de Espanhol da Universidade de Valladolid*, que teve início em 2012, quando Fuertes-Olivera e Bergenholtz firmaram um contrato com a empresa dinamarquesa *Ordbogen.com* que, além de ser especializada em desenvolvimento de sistemas computacionais para dicionários, se encarrega de administrar a comercialização do produto final (FUERTES-OLIVEIRA; BERGENHOLTZ, 2015, p. 73).

Diante do exposto, no intuito de exemplificar o uso de soluções computacionais inovadoras no labor lexicográfico e dialetal, dispensando a contratação de serviços terceirizados de programação, o tópico a seguir descreve a metodologia aplicada na construção de ferramentas computacionais personalizadas, destinadas a atender necessidades de manipulação de dados dialetais e lexicográficos.

3 Metodologia

O *corpus* utilizado como fonte de dados para este artigo advém do Atlas Linguístico do Brasil (COMITÊ NACIONAL DO PROJETO ALIB, 2001), cujo objetivo principal é identificar e descrever o falar característico em todos os estados da Federação distribuídos pelas cinco regiões do Brasil.

Para tanto, o Projeto ALiB iniciou, em 2001, a árdua tarefa de coletar a fala de entrevistados que atendessem ao perfil definido de acordo com os pressupostos da Geolinguística e os objetivos do atlas a ser produzido. Assim, a coleta de dados foi realizada em 250 localidades[19] distribuídas por todo o Brasil por meio de entrevistas gravadas em áudio. Essa coleta de dados foi finalizada em 2013 e resultou na formação de um robusto *corpus* oral de dados dialetais que tem possibilitado a cartografia de dados linguísticos e estudos sobre o conteúdo das cartas linguísticas, publicadas no volume 2 do ALiB (CARDOSO et al., 2014), além de subsidiar diversificados estudos de ordem lexical, semântico, fonético-fonológico, sintático e meta(lexical)[20].

Destaca-se, no montante dos trabalhos realizados a partir do corpus do ALiB, aqueles que se dedicaram ao tratamento lexicográfico aos dados orais e, desse modo, elaboraram produtos lexicográficos, a saber: o *Vocabulário Dialetal Baiano* (NEIVA, 2017); o *Vocabulário Dialetal do Centro-Oeste: interfaces entre a Lexicografia e a Dialetologia* (COSTA, 2018); o *Vocabulário dialetal da região Norte*

[18] "[...] uma função lexicográfica pode ser definida como a assistência fornecida por uma obra lexicográfica para satisfazer tipos específicos de necessidades de informação pontual que um tipo específico de usuário potencial possa ter em um tipo específico de situação extralexical. A referida assistência é conseguida através de dados lexicográficos cuidadosamente elaborados e disponibilizados para consulta" (TARP, 2015, p. 36).

[19] A rede de pontos do Projeto ALiB reúne 250 localidades, 225 localidades do interior do Brasil e 25 capitais distribuídas pelas cinco regiões do país. Maiores informações em: https://alib.ufba.br/content/rede-de-pontos. Acesso em: 10 out. 2022.

[20] Os trabalhos realizados e em andamento a partir dos dados do ALiB podem ser encontrados no site do Projeto: https://alib.ufba.br/. Acesso em: 10 out. 2022.

do Brasil: um estudo das capitais com base nos dados do Projeto ALiB (CORREIA DE SOUSA, 2019) e o *Vocabulário Dialetal Maranhense: a contribuição do Maranhão para o Dicionário Dialetal Brasileiro* (MARAMALDO FERREIRA, 2019).

Ressalta-se, ainda, que os trabalhos supracitados, além da tese de Doutorado, em andamento, que forneceu os dados para as discussões deste artigo, estão filiados ao Projeto Dicionário Dialetal Brasileiro (DDB)[21] (MACHADO FILHO, 2010) que tem como objetivo dar tratamento lexicográfico ao corpus dialetal do Projeto ALiB.

Para evitar ambiguidades, assume-se, neste artigo, que o termo *ferramentas computacionais* está sendo utilizado para denominar as linhas de código escritas para realizar tarefas específicas a partir de um ambiente informatizado. Isso significa que essas ferramentas foram desenvolvidas com o uso de linguagens de programação e de marcação para atender, exclusivamente, as demandas da pesquisa. Destaca-se, ainda, que tais ferramentas são de código aberto e livre, podendo ser utilizadas e/ou modificadas em outros projetos.

O ambiente utilizado para gerenciar o banco de dados em *XML* por meio das ferramentas aqui apresentadas é o *software BaseX*[22], também de código aberto. Essa plataforma permite a visualização dos dados em diferentes formatos e contém um editor de texto em que é possível escrever as linhas responsáveis por recuperar informações do *XML*. Além disso, o *BaseX* oferece a possibilidade de desenvolver um *website* integrado ao banco de dados, além de permitir a visualização do projeto em uma aba do navegador de internet por meio da abertura de uma porta local.

Desse modo, o tratamento eletrônico dispensado aos dados do ALiB documentados nas 18 localidades do interior da região Norte do Brasil, se caracteriza pela organização do *corpus* dialetal de maneira lexicográfica em um banco de dados em *XML*. Assim, o arquivo com extensão *.xml*[23] foi escrito levando em consideração as variáveis dialetais presentes no corpus do ALiB e os elementos previstos para compor a microestrutura do protótipo do VoDiNorte, conforme o detalhamento a seguir.

4 Apresentação dos dados

Tendo em vista que o perfil dos informantes no Projeto ALiB considera as variáveis sexo, idade, escolaridade e localidade e que a microestrutura do protótipo do VoDiNorte é composta por onze elementos – lema, classe gramatical, variação fonética, definição, exemplo, informante, áudio, pergunta, representação cartográfica e remissiva –, a estrutura do *XML* foi escrita de modo a contemplar todas essas informações como detalhado na Lista 1:

Lista 1. Estrutura do arquivo *XML*,

```
1  <entrada id="fauna.faun.insetos.1064" abc="m">
2    <lema>mosquito</lema>
3    <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>
4    <ex>Carapanã, musquito. (É a mesma coisa carapanã i musquito?) É a mesma
         coisa. (É o mesmo bichinho?) É o mesmo bichinho.</ex>
5    <obs></obs>
6   <fone>musquito</fone>
7    <aud src="fala-id-1064.mp3" type="mp3">1:05:15</aud>
8    <ver name="carapanã" ref="fauna.faun.insetos.1534"/>
9    <info sexo="Masculino" escolaridade="fundamental" idade="jovem" >29 anos<
         /info>
10   <lg ponto="4" cidade="São Gabriel da Cachoeira" estado="AM"/>
11   <gram>Substantivo masculino</gram>
12   <def>Inseto pequeno que voa e pica, semelhante ao carapanã.</def>
13   <map src="mapa-1064.jpg" type="jpg"/>
```

[21] Coordenado pelo professor Américo Venâncio Machado Filho da Universidade Federal da Bahia.

[22] O ambiente de desenvolvimento utilizado foi o *software BaseX* que permite, entre outras funções, recuperar dados escritos em *XML* por meio de linhas de código escritas em seu editor. O *software* pode ser acessado em: https://www.basex.org. Acesso em: 12 nov. 2022.

[23] Arquivos com extensão *.xml* podem ser escritos em editores simples como, por exemplo, o bloco de notas do computador, ou por uma variedade de editores de texto elaborados para facilitar o trabalho dos programadores. O banco de dados do VoDiNorte foi escrito com o uso do *software jEdit* https://www.jedit.org. Acesso em: 14 set. 2022.

```
14  </entrada>
```
Fonte: Elaboração do autor.

Observa-se, a partir da Lista 1, que as informações foram armazenadas em campos distintos de acordo com cada uma das 202 perguntas e respostas do Questionário Semântico-lexical (QSL) do Projeto ALiB. Esses conteúdos foram alocados em campos específicos que, na linguagem de marcação *XML*, denominam-se *tags*[24]. Desse modo, as linhas 1 a 14 indicam a escrita de 14 *tags*, a saber:

- Linha 1: Formada pela *tag* de abertura `<entrada>` e possui dois atributos em que o primeiro é uma identificação (`id="fauna.faun.insetos.1064"`) escrita a partir de uma sequencia de caracteres que orientam uma pesquisa por uma área semântica no protótipo do VoDiNorte, seguido de uma sequencia numérica que é única no *XML*. O segundo atributo é a letra inicial do lema (`abc="m"`). Vale acrescentar que o fechamento dessa *tag* está na linha 14 configurando, assim, um bloco de dados;
- Linha 2: Indica a existência de uma *tag* de abertura `<lema>` e de fechamento `</lema>` que armazena o lema em questão (mosquito);
- Linha 3: Elaborada para armazenar os dados relacionados à pergunta do QSL possui o atributo *campo*, que indica qual das 12 áreas semânticas[25] a pergunta pertence (`campo="Fauna"`) e ainda outro atributo que indica o número da referida pergunta (`ref="QSL-88/ALiB"`). Além disso, antes do fechamento dessa *tag* a pergunta foi escrita por extenso;
- Linha 4: Destinada a exibir a fala mencionada pelo informante no momento em que responde à pergunta 88, sendo considerada como um exemplo de uso e representada pela tag de abertura <ex> e pela tag de fechamento </ex>.;
- Linha 5: Representa a formação de uma *tag* criada para registrar observações que o pesquisador julgar necessárias no momento em que está ouvindo as entrevistas e transcrevendo os dados no *XML*;
- Linha 6: Reservada para indicar a *tag* de abertura `<fone>` e de fechamento `</fone>` responsável por armazenar a variação fonética do entrevistado, ou seja, o modo como foi pronunciado o lema em questão;
- Linha 7: Formada por uma *tag* que contém dois atributos que especificam os dados utilizados pela ferramenta de áudio do VoDiNorte, tendo em vista que o protótipo pode reproduzir um trecho de fala dos informantes em alguns verbetes. Assim, o primeiro atributo armazena o nome do arquivo de áudio que deverá ser executado (`src="fala-id-1064.mp3"`) e o segundo atributo especifica o tipo de arquivo (`type="mp3"`). Há, ainda, uma indicação da hora, minuto e segundo em que a fala do informante aparece no arquivo de áudio que contém toda a entrevista;
- Linha 8: Mostra o funcionamento da *tag* responsável pelo sistema de remissivas do protótipo do VoDiNorte e possui dois atributos em que o primeiro faz referência ao nome da entrada que está relacionada ao verbete mosquito (`name"=carapanã"`) e o segundo apresenta a *id* do lema *carapanã*, que funciona como um endereço para recuperar uma entrada por meio da ferramenta de remissivas;
- Linha 9: Especifica os dados do informante por meio de três atributos, isto é, um para o sexo (`sexo ="Masculino"`), outro para o nível de escolaridade (`escolaridade="fundamental"`) e o terceiro para a idade (`idade="jovem"`). Além disso, há uma informação que exibe a idade do entrevistado antes do fechamento da tag `</info>`;
- Linha 10: Composta por três atributos é uma *tag* que armazena os dados relacionados à localidade, indicando o número da rede de pontos do Projeto ALiB (`ponto""=4`), bem como a cidade (`cidade = "São Gabriel da "Cachoeira`) e o estado (`estado"="AM`);
- Linha 11: Faz referência à classe gramatical do verbete por meio de uma *tag* do tipo elemento;
- Linha 12: Armazena a definição do verbete a partir de uma *tag* do tipo elemento;

[24] As tags em *XML* podem assumir dois modelos de apresentação, a saber: as do tipo elemento destinadas a informações textuais, em geral, extensas; e as do tipo atributo, que foram utilizadas na pesquisa para agregar particularidades do texto. Não há regras fixas para essas composições, de modo que cada projeto deve analisar e testar a estrutura de *tags* que melhor atender aos objetivos de uma determinada pesquisa.

[25] As 202 perguntas do QSL/ALiB estão organizadas a partir de 12 áreas semânticas, a saber: acidentes geográficos, fenômenos atmosféricos, astros e tempo, atividades agropastoris, fauna, corpo humano, ciclos da vida, convívio e comportamento social, religião e crenças, jogos e diversões infantis, habitação, alimentação e cozinha, vestuário e acessórios, vida urbana.

- Linha 13: Apresenta o nome do arquivo de imagem (`src="mapa-1064.jpg"`) que deve ser recuperado pela ferramenta de representação cartográfica do protótipo do VoDiNorte, bem como indica o tipo de arquivo utilizado (`type"="jpg`);
- Linha 14: Destinada ao fechamento desse bloco de dados por meio da *tag* de fechamento `</entrada>`.

Diante do exposto é possível observar que a linguagem de marcação *XML* funciona de modo a rotular uma variedade de informações para, posteriormente, permitir o gerenciamento desses dados a partir de ferramentas computacionais. Assim, o esquema da Lista 1 pode ser aludido a uma "gaveta" localizada em um grande armário. Dentro dessa "gaveta" há subdivisões para agrupar os dados por categorias que estão armazenadas em *tags*.

Nessa analogia, o processo de armazenamento de informações no banco de dados segue a dinâmica de abrir uma "gaveta" e alocar cada tipo de conteúdo em uma subdivisão específica. Para cada conjunto de dados uma nova "gaveta" é aberta e assim, sucessivamente, formando um conjunto de centenas, de milhares de "gavetas". Esse procedimento garante a formação de um repositório de dados estruturados, ou seja, padronizados com o auxílio de um *document type definition (DTD)*[26]. Esse formato estruturado é uma maneira de viabilizar a manipulação eletrônica mediante ferramentas computacionais e que podem, a partir do uso de linguagens de programação, alimentar o projeto de um *website*.

Vale ressaltar, ainda, que as *tags*[27] receberam nomes curtos e, em alguns casos, uma forma abreviada foi utilizada para melhorar o aspecto visual do banco de dados, facilitando a identificação das informações pelo olhar humano.

Após a estruturação do banco de dados, seguiu-se com o armazenamento dos dados em cada *tag*. Esse procedimento continua em andamento, pois a quantidade de entrevistas a serem ouvidas e transcritas é expressiva. No entanto, é possível avançar para a próxima etapa a partir dos dados armazenados até o momento, isto é, a escrita das ferramentas computacionais responsáveis pela recuperação de informações.

É importante destacar que essas ferramentas foram construídas por meio de comandos escritos em uma linguagem[28] que o computador possa compreender e são executados no editor do *software BaseX*. Desse modo, as possibilidades de execução de tarefas de cada ferramenta dependem, exclusivamente, de dois fatores, a saber: i) a estrutura de cada *tag* do *XML*; ii) a escrita correta das instruções na linguagem de consulta *X-Query*.

Desse modo, para visualizar os dados armazenados na *tag* `<lema></lema>` e na *tag* `<ex></ex>`, apresentadas na Lista 1, deve-se orientar a máquina com as seguintes instruções:

Lista 2. Instruções para recuperação de dados de uma entrada específica.

```
1  for $x in db:open ("corpus-oral-1")//entrada
2     where $x//@id="fauna.faun.insetos.1064"
3  return ($x//lema,$x//ex)
```

Fonte: Elaboração do autor.

Observa-se que a Lista 2 é composta por três linhas de código. A primeira cria a variável $x – que pode ser aludida a uma caixa virtual – que acessa o banco de dados corpus-oral-1 e, por sua vez, todas as *tags* `<entrada></entrada>`. A linha 2 cria um filtro que solicita a exibição do conteúdo armazenado na *tag* `</entrada></entrada>` que possui o seguinte nome de identificação: `fauna.faun.insetos.1064`. Em seguida, a linha 3 orienta o *software* para mostrar apenas os conteúdos da *tag* `<lema></lema>` e `<ex></ex>` processados dentro da variável $x (caixa virtual). O resultado desse

[26] O documento com extensão *.dtd* é utilizado para escrever os parâmetros de funcionamento do arquivo *XML*. O *DTD* é anexado ao *XML* por meio de uma declaração escrita no início do arquivo e sua função é garantir que o a estrutura das *tags* permaneçam do modo que foram planejadas. Assim, caso algum caractere de qualquer *tag* for apagado, acidentalmente, uma mensagem de erro será emitida indicando ao usuário a linha e o tipo de dado que se encontra em desacordo com o *DTD*.

[27] A nomeação de *tags* em um arquivo *.xml* é livre, ou seja, o pesquisador pode escolher qualquer palavra ou conjunto de caracteres que faça mais sentido do ponto de vista humano, pois para o computador o que importa é que existem dados etiquetados sob determinada *tag* que pode ser acessada mediante um nome.

[28] A linguagem utilizada no projeto para dar instruções à máquina, no editor do *BaseX*, denomina-se *X-Query*. Trata-se de uma linguagem de consulta que utiliza elementos de programação para recuperar dados em formato *XML*.

conjunto de instruções é o seguinte:

Lista 3. Resultados das instruções escritas na Lista 2

```
1  <lema>mosquito</lema>
2  <ex>Carapanã, musquito. (É a mesma coisa carapanã i musquito?) É a mesma
       coisa. (É o mesmo bichinho?) É o mesmo bichinho.</ex>
```

Fonte: Elaboração do autor.

Constata-se, por meio da Lista 3, que o resultado devolvido ao usuário no momento em que se enviou o conjunto das instruções ao computador (Lista 2) fecha um ciclo de trabalho em que houve a solicitação de uma tarefa e sua execução. Esse processo pode ser compreendido como um pequeno[29] programa, ou seja, uma ferramenta computacional destinada a realizar um trabalho específico. Partindo desse entendimento, outras ferramentas podem ser escritas com a finalidade de realizar tarefas variadas, ou seja, para cada recuperação de informação que se queira realizar no banco de dados uma pequena ferramenta deve ser criada.

Para exemplificar essas possibilidades, três questões foram formuladas cujas respostas podem ser retiradas do banco de dados, a saber: i) Qual foi a variação fonética ocorrida nas respostas das informantes mulheres na cidade de Tefé? ii) Quais foram as respostas fornecidas pelos idosos, do sexo masculino, para a pergunta 12[30] do QSL-ALiB no interior do estado do Amazonas[31]? iii) Como recuperar determinada unidade lexical nas falas dos informantes? Essas questões estão detalhadas a seguir.

4.1 Qual foi a variação fonética ocorrida nas respostas das informantes mulheres na cidade de Tefé/AM?

Para responder a essa questão foram escritas as seguintes linhas no editor do *BaseX*:

Lista 4. Instruções para a recuperação de dados referentes à variação fonética.

```
1  for $x in db:open ("corpus-oral-1")//entrada
2    where $x//@sexo="F" and $x//@cidade="Tefé"
3  return ($x//@id,$x//lema,$x//fone,$x//perg[@ref],$x//@cidade)
```

Fonte: Elaboração do autor.

Como se pode observar na Lista 4, é preciso orientar o computador sobre a localização do arquivo *.xml* (linha 1) e, posteriormente, escrever qual tipo de filtro a máquina deve realizar, ou seja, indicar que se deseja recuperar dados de informantes mulheres (`$x//@sexo="F"`) e moradoras do município de Tefé (`$x//@cidade="Tefé"`). Para finalizar essa solicitação, na linha 3, após o comando *return* especificam-se quais elementos da microestrutura se deseja visualizar. Neste caso, após cada $x – a caixa virtual – indica-se a *tag* que se quer visualizar como resultado, ou seja, as *tags* que correspondem à identificação da entrada (`id`), ao lema, à variação fonética, à pergunta e à cidade. O resultado dessas instruções pode ser visto na Lista 5:

Lista 5. Resultados das instruções escritas na Lista 4

```
1   id="acid.geo.água.3137"
2  <lema>igarapé</lema>
3  <fone>garapé</fone>
4  <perg campo="Acidentes geográficos" ref="QSL-1"/>
5   cidade="Tefé"
6   ...
```

Fonte: Elaboração do autor.

Como os resultados exibidos por esta solicitação são extensos, a Lista 5 mostra apenas o primeiro conjunto de dados o qual está descriminado a id do lema em questão (linha 1), o lema (linha 2), a

[29] O termo pequeno, nessa afirmativa, alude a um programa de baixa complexidade, ou seja, simples de ser escrito e que pode ser desenvolvido por iniciantes.
[30] Existem outros nomes para temporal? QSL-12.
[31] A rede de pontos do interior do Projeto ALiB, do estado do Amazonas, reúne quatro localidades: São Gabriel da Cachoeira, Tefé, Benjamin Constant e Humaitá.

variação fonética (linha 3), a área semântica e a pergunta do QSL-ALiB correspondente (linha 4), além da localidade (linha 5).

Acrescenta-se, ainda, que a escrita dessas ferramentas tende a ser facilitada ao pesquisador após o desenvolvimento do primeiro pequeno programa, pois a estrutura da primeira linha, exibida na Lista 4, se manterá igual. Dessa forma, o que deverá ser modificado são os caminhos para o acesso dos conteúdos das *tags* que são escritos nos comandos das linhas subsequentes como, por exemplo, se ilustra no caso da próxima pergunta.

4.2 Quais foram as respostas dadas pelos idosos, do sexo masculino, para a pergunta 12 do QSL-ALiB no interior do estado do Amazonas?

Na construção dessa ferramenta as linhas 2 e 3 foram escritas da seguinte forma, tendo em vista não haver necessidade de alterar as instruções da linha 1:

Lista 6. Instruções para a recuperação de dados filtrados pelas variantes *idade, sexo, pergunta* e *localidade*.

```
1  for $x in db:open ("corpus-oral-1")//entrada
2    where $x//@sexo="M" and $x//@idade="I" and $x//perg[@ref="QSL-12"] and $x
       //@estado="AM"
3    return ($x//lema,$x//@cidade)
```

Fonte: Elaboração do autor.

Os comandos escritos na linha 2, apresentados na Lista 6, orientam o *software* para acessar os dados referentes ao sexo masculino (`$x//@sexo="M"`), à idade (`$x//@idade="I"`), à pergunta 12 do QSL-ALiB (`$x//perg[@ref="QSL-12"]`) e ao estado do Amazonas (`$x//@estado="AM"`). Cada solicitação é interligada pelo comando *and* e, como resultado, a máquina deverá exibir apenas o lema e a cidade, conforme as instruções da linha 3. O resultado desse conjunto de códigos pode ser visto na Lista 7:

Lista 7. Resultados das instruções escritas na Lista 6

```
1  <lema>tempestade</lema>
2  cidade="São Gabriel da Cachoeira"
3  <lema>ventania</lema>
4  cidade="São Gabriel da Cachoeira"
5  <lema>tempestade</lema>
6  cidade="Benjamin Constant"
7  <lema>temporal forte</lema>
8  cidade="Tefé"
9  <lema>temporal</lema>
10 cidade="Humaitá"
11 <lema>chuva grossa</lema>
12 cidade="Humaitá"
```

Fonte: Elaboração do autor.

Percebe-se, por meio da Lista 7, que os resultados foram mostrados linha a linha, isto é, em uma sequência que foi especificada na linha 3, da Lista 6. Como em cada município, dos quatro informantes entrevistados apenas um se enquadra no perfil masculino e idoso, as cidades que aparecem duplicadas nos resultados indicam se tratar de respostas diferentes do mesmo informante. É importante frisar que, caso o pesquisador deseje visualizar outros resultados além dos exibidos na Lista 7, poderá orientar o sistema, na linha 3, da Lista 6, para que exiba dados de outras *tags*.

Acrescenta-se, ainda, que a pergunta que motivou este tópico permite que o investigador realize um levantamento, a fim de comparar os dados de homens *versus* mulheres e jovens *versus* idosos em todas as localidades ou em um município específico escrevendo, desse modo, novas ferramentas como o ilustrado a partir da pergunta seguinte.

4.3 Como recuperar determinada unidade lexical nas falas dos informantes?

Esta pergunta foi respondida por uma ferramenta computacional criada para acessar todas as *tags* *<ex></ex>* do banco de dados. Nesse exemplo, escolheu-se buscar a unidade lexical *balo* que

chamou a atenção devido ao seu emprego e a sua própria composição lexical. A escrita da ferramenta está ilustrada na Lista 8:

Lista 8. Instruções para a recuperação de dados a partir de uma unidade lexical.

```
1  for $x in db:open ("corpus-oral-1")//entrada
2    where $x//ex[text() contains text{"balo"}]
3    return ($x//lema,$x//ex,$x//@sexo,$x//@idade,$x//@escolaridade,$x//
         @cidade,$x//@estado)
```

Fonte: Elaboração do autor.

Verifica-se, por meio da Lista 8, que a linha 1 permanece inalterada e que na linha 2 há um comando que especifica a *tag* e o conteúdo de texto que deverá ser recuperado, escrito entre aspas e dentro das chaves. Por sua vez, a linha 3 orienta o computador a exibir nos resultados os conteúdos das *tags* que dizem respeito ao lema, exemplo, sexo, idade, escolaridade, cidade e estado. Esses resultados são descritos na Lista 9:

Lista 9. Resultados das instruções escritas na Lista 8

```
1  <lema>baladeira</lema>
2  <ex>Baladera. Baladera. Eu balo us cachorro às vezis. Quando num mi deixa
       durmi. Aí eu saiu i eu balo. Eu tenho uma baladera aí. [risos] Acho qui
        tá proibido, né? (Ah, é?) Acho que proibiu. Federal proibiu. Qui num
       tem qui vendê nu supermercado. (Mas tá na casa da genti, né...) Ah, us
       cachorro num tão mi deixando durmi. Aí eu balo us cachorro.</ex>
3  sexo="F"
4  idade="I"
5  escolaridade="F"
6  cidade="São Gabriel da Cachoeira"
7  estado="AM"
```

Fonte: Elaboração do autor.

É possível identificar, a partir da Lista 9, que foi o lema *baladeira* (linha 1) que motivou a menção da unidade lexical *balo* (linha 2) pela informante do sexo feminino (linha 3), idosa (linha 4), com nível fundamental de escolaridade (linha 5) e moradora da cidade de São Gabriel da Cachoeira (linha 6), estado do Amazonas (linha 7).

Destaca-se, ainda, que essa ferramenta computacional pode pesquisar outros itens lexicais que estejam presente nas *tags* `<ex></ex>`, bastando substituir a unidade lexical *balo* da linha 2, Lista 8, por outro item que se deseja investigar. Assim, a cada modificação feita na estrutura dos códigos uma nova ferramenta computacional é produzida.

5 Discussão dos resultados

Com base na experiência vivenciada no decorrer da produção das ferramentas computacionais, aqui apresentadas, constatou-se que o método utilizado é eficiente para recuperar dados estruturados em *XML*. Porém, para que se tenha êxito nessa tarefa o estudioso deve lançar mão de conhecimentos pontuais relacionados à construção de bancos de dados em *XML*, bem como saber gerenciar as informações no *BaseX* por meio da linguagem de consulta *X-Query*.

Destaca-se, dessa maneira, que o uso dessas ferramentas possibilitam ao pesquisador um estudo sistematizado de um corpus em diferentes perspectivas, pois é possível recuperar informações a partir da visualização dos conteúdos armazenados nas *tags* do *XML* e, ainda, refinar essa busca acrescentando mais opções de filtragem de dados por meio do comando *and*, conforme foi ilustrado na Lista 6.

Partindo dessas possibilidades de recuperação de informação um linguista poderá, por exemplo, desenvolver estudos de ordem semântica, sintática, fonética e fonológica. Para tanto, deverá escrever uma expressão *X-Query* que atenda a sua demanda de pesquisa e, nesse ponto, o método de consulta apresentado neste artigo se mostra articulado, pois pode ser ligeiramente modificado para atender a novos tipos de recuperação de dados.

No entanto, tais ferramentas são limitadas a exibir ao usuário informações que podem ser recuperadas a partir das possibilidades existentes dentro da escrita de uma expressão *X-Query*, além de

estarem condicionadas à forma com que o arquivo *XML* foi planejado, ou seja, que tipo de dados foram armazenados em *tags* do tipo atributo e quais informações foram alocadas em *tags* do tipo elemento, caracterizando um tipo de estrutura *XML* adequada para cada tipo de projeto.

Isso significa que o planejamento prévio da composição das *tags* do *XML* deve estar em conformidade com o tipo de recuperação de informação que se deseja fazer. Asim, antes de escrever a estrutura arbórea do banco de dados é preciso se perguntar que tipo de operações de recuperação de informação o sistema deverá executar.

De posse dessa resposta o pesquisador poderá rascunhar as *tags* que irão compor o banco de dados e realizar testes com uma amostra de dados para verificar se a estrutura do *XML* corresponde às necessidades do projeto. Recomenda-se, dessa forma, só iniciar o armazenamento dos dados, definitivamente, após a conclusão da etapa de testes.

Em suma, além de apresentar as possibilidades de manipulação de dados dialetais e lexicográficos este artigo também tem a finalidade de mostrar que é possível dar tratamento informatizado a um *corpus* de estudo sem ser um especialista em programação. Assim, espera-se que a experiência compartilhada neste estudo possa motivar outros pesquisadores a se enveredarem pelo mundo das linguagens de programação, a fim de construírem suas próprias ferramentas computacionais. Para tanto, a metodologia apresentada neste trabalho pode ser sintetizada em duas etapas principais, a saber:

i Criar um banco de dados em *XML*, estruturado segundo os objetivos de cada pesquisa;

ii Utilizar a linguagem de consulta *X-Query*, no editor do *software BaseX*, para manipular os dados.

Evidentemente, como já mencionado, o pesquisador deverá investir na aprendizagem de conteúdos específicos relacionados ao *XML, X-Query* e *BaseX* e, em sua jornada, outros temas poderão ser incluídos no roteiro de conteúdos a serem desbravados. No entanto, a investida é recompensadora, pois, além de automatizar tarefas por meio de linhas de código permite a utilização do banco de dados em *XML* em outros projetos, tendo em vista que sua linguagem é compatível com outros sistemas informáticos.

6 Considerações finais

As soluções informáticas demostradas neste artigo foram criadas a partir de demandas específicas. Nesse sentido, vale destacar a liberdade e a autonomia do pesquisador para editar as linhas de código sempre que necessário, a fim de adaptar as ferramentas para o seu uso, atentando, apenas, para as possibilidades que podem ser traçadas entre as *tags* do *XML* e os comandos da linguagem *X-Query*.

Destaca-se, ainda, que os conhecimentos a serem buscados pelo investigador, no intuito de não terceirizar a mão de obra relacionada à programação, dependerão do projeto a ser executado. Nesse sentido, é importante procurar uma consultoria com um profissional da área, pois somente um especialista poderá sugerir as possibilidades existentes no abrangente universo da computação, para planejar um roteiro de estudos com os conteúdos que deverão ser aprendidos.

Outro ponto importante a ser considerado é que a recuperação de dados a partir de soluções informatizadas não oferece apenas rapidez de acesso à informação. O uso de ferramentas computacionais personalizadas, na pesquisa científica, amplia os horizontes de observação, aumentando as possibilidades de análise e reflexão. Desse modo, o estudioso pode utilizar-se de técnicas de PLN para observar a combinação das unidades lexicais em uma sentença, por meio de dados probabilísticos. Nessa abordagem o linguista pode compreender o princípio do funcionamento dos assistentes de escrita que sugerem ao usuário opções de palavras que se enquadram na sentença, a partir de dados da frequência de combinações sintáticas em uma língua natural.

Finalmente, a compatibilidade dos dados é uma característica de um trabalho desta natureza e, desse modo, destaca-se que a combinação de linguagens de programação com um *corpus* estruturado em *XML* potencializa o surgimento de estudos e produtos futuros.

Referências

CARDOSO, Suzana Alice Marcelino. A dialetologia e os estudos da variação linguística. In: CARDOSO, Suzana Alice Marcelino (Ed.). *Geolinguística - tradição e modernidade*. São Paulo: Parábola Editorial, 2010. P. 15–30.

CARDOSO, Suzana Alice Marcelino et al. *Atlas linguístico do Brasil: Cartas Linguísticas 1*. Londrina: EDUEL, 2014. v. 2.

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.

CORREIA DE SOUSA, Cemary. *Vocabulário dialetal da região norte do Brasil: um estudo das capitais com base nos dados do projeto ALIB*. 2019. 134 f. Mestrado em Língua e Cultura – Universidade Federal da Bahia, Salvador.

COSTA, Daniela de Souza Silva. *Vocabulário Dialetal do Centro-Oeste: interfaces entre a Lexicografia e a Dialetologia*. 2018. 353 f. Doutorado em Estudos da Linguagem – Universidade Estadual de Londrina, Londrina.

FUERTES-OLIVEIRA, Pedro Antonio; BERGENHOLTZ, Henning. Introduction: The Construction of Internet Dictionaries. In: FUERTES-OLIVEIRA, Pedro Antonio; BERGENHOLTZ, Henning (Ed.). *e-Lexicography: The Internet, Digital Initiative and Lexicography*. London/New York: Continuum, 2011. P. 1–16.

FUERTES-OLIVEIRA, Pedro Antonio; BERGENHOLTZ, Henning. Los Diccionarios en Línea de Español "Universidad de Valladolid." *Estudios de Lexicografía. Revista Mensual del grupo de las dos vidas de las palabras*, n. 4, p. 71–98, jun. 2015. Disponível em: <https://issuu.com/ldvp/docs/elex_4-_def>. Acesso em: 2 ago. 2022.

KEDIA, Aman; RASU, Mayank. *Hands-on Python natural language processing: explore tools and techniques to analyze and process text with a view to building real-world NLP applications*. Birmingham: Packt Publishing Ltd, 2020.

LEROYER, Patrick. Change of paradigm: from Linguistics to Information Science and from dictionaries to lexicographic information tools. In: FUERTES-OLIVEIRA, Pedro Antonio; BERGENHOLTZ, Henning (Ed.). *e-Lexicography: The Internet, Digital Initiative and Lexicography*. London/New York: Continuum, 2011. P. 121–140.

MACHADO FILHO, Américo Venâncio Lopes. Um ponto de interseção para a dialectologia e a lexicografia: a proposição de um dicionário dialetal brasileiro com base nos dados do ALiB. *Estudos Linguísticos e Literários*, v. 41, p. 49–70, 2010.

MANNING, Christopher D. *Introduction to information retrieval*. Cambridge: Cambridge University Press, 2008.

MANNING, Christopher D; SCHUTZE, Hinrich. *Foundations of statistical natural language processing*. Cambridge: MIT press, 1999.

MARAMALDO FERREIRA, Camila. *Vocabulário Dialetal Maranhense: a contribuição do Maranhão para o Dicionário Dialetal Brasileiro 2019*. 2019. 119 f. Mestrado em Letras – Universidade Federal do Maranhão, São Luís.

NEIVA, Isamar. *Vocabulário Dialetal Baiano*. 2017. 270 f. Doutorado em Língua e Cultura – Universidade Federal da Bahia, Salvador.

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.

RADTKE, Edgar; THUN, Harald. Nuevos caminos de la geolinguística románica. In: RADTKE, Edgar; THUN, Harald (Ed.). *Neue Wege der Romanischen Geolinguistik*. Kiel: Westensee-Verlag, 1996. P. 25–49.

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. *Lexicography in the borderland between knowledge and non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Niemeyer, 2008.

TARP, Sven. Lexicographical and other e-tools for consultation purposes: towards the individualization of needs satisfaction. In: FUERTES-OLIVEIRA, Pedro Antonio; BERGENHOLTZ, Henning (Ed.). *e-Lexicography: The Internet, Digital Initiative and Lexicography*. London/New York: Continuum, 2011. P. 54–70.

TARP, Sven. La teoría funcional en pocas palabras. *Estudios de Lexicografía. Revista Mensual del grupo de las dos vidas de las palabras*, v. 4, p. 31–42, 2015. Disponível em: <https://issuu.com/ldvp/docs/elex_4-_def>. Acesso em: 2 ago. 2022.