DOI: 10.14393/DL52-v16n4a2022-11

A construção de um banco de dados lexicográficos em XML a partir de dados dialetais: o Processamento Automático de Linguagem Natural (PLN)

The construction of a lexicographic database in XML from dialectal data: the Natural Language Processing (NLP)

Jorge Luiz Nunes dos SANTOS JUNIOR[*]
Aparecida Negri ISQUERDO[**]

RESUMO: Este artigo situa-se na interface entre a Lexicografia (PORTO DAPENA, 2002; HARTMANN, 2016), a Dialetologia (CARDOSO, 2010; CHAMBERS; THUDGILL, 1994) e a Linguística Computacional (HABERT, 2004; PÉREZ HERNÁNDEZ; MORENO ORTIZ, 2009; HAUSSER, 2014; KURDI, 2016). Objetiva-se discutir a proposta de construção de um banco de dados em *XML (Extensible Markup Language)*, explorando os resultados obtidos com o PLN (Processamento Automático de Linguagem Natural). O arquivo *XML* também se fundamenta em parâmetros da Lexicografia Dialetal (EZQUERRA, 1997; NAVARRO CARRASCO, 1993) e está sendo alimentado com dados dialetais oriundos do Projeto Atlas Linguístico do Brasil (ALiB) documentados na região Norte do país. Para tanto, utilizou-se como editor de

ABSTRACT: This paper is situed at the interface between Lexicography (PORTO DAPENA, 2002; HARTMANN, 2016), Dialectology (CARDOSO, 2010; CHAMBERS; THUDGILL, 1994) and Computational Linguistics (HABERT, 2004; PÉREZ HERNÁNDEZ; MORENO ORTIZ, 2009; HAUSSER, 2014; KURDI, 2016). The objective is to discuss the proposal of building a database in XML (Extensible Markup Language), exploring the results obtained with NLP (Natural Language Processing). The XML file is also based on parameters of Dialectal Lexicography (ESQUERRA, 1997; NAVARRO CARRASCO, 1993) and is being fed with dialectal data from the project Atlas Linguístico do Brasil (ALiB) documented in the country's Northern region. Therefore, the jEdit software was used as a text editor and, to manage the database, the BaseX program. The

[*] Doutorando do Programa de Pós-Graduação em Letras da Universidade Federal de Mato Grosso do Sul, *campus* de Três Lagoas (UFMS/CPTL). Bolsista CAPES. ORCID: https://orcid.org/0000-0002-1111-4148. jorgesantosjunior@gmail.com.

[**] Doutora em Letras (Linguística e Língua Portuguesa) pela UNESP/Araraquara. Pesquisadora Sênior/UFMS. Docente da UFMS – Estudos de Linguagens/FAALC e Letras/CPTL. ORCID: https://orcid.org/0000-0003-1129-5775. aparecida.isquerdo@gmail.com.

texto o software *jEdit* e, para gerenciar o banco de dados, o programa *BaseX*. A extração das informações linguísticas foi realizada, no *BaseX*, a partir de uma amostra de dados e com o auxílio de expressões *X-Query*. Assim, foram executadas as seguintes manipulações de dados: i) localização de uma unidade lexical específica; ii) visualização de qualquer dado da microestrutura filtrada pelas variáveis sexo, idade, escolaridade e localidade; iii) seleção de informações a partir de uma das 14 áreas semânticas em que as questões do questionário semântico-lexical do ALiB foram organizadas. Em síntese, entende-se que a construção do banco de dados em *XML* confere agilidade em relação à extração de informações e compatibilidade dos dados para executar interfaces com outras aplicações como, por exemplo, a elaboração de um produto lexicográfico a ser publicado em suporte *on-line*.

PALAVRAS-CHAVE: Lexicografia Dialetal. Linguística Computacional. Banco de dados em XML. PLN.

linguistic information extraction was performed in the BaseX, from a sample of data and with the X-Query expressions support. Thus, the following data manipulations were performed: i) location of a specific lexical unit; ii) visualization of any microstructure data filtered by variables gender, age, education and location; iii) selection of information from one of the 14 semantic areas in which the questions of the ALiB semantic-lexical questionnaire were organized. In summary, it is understands that the construction of a XML database provides agility in concerning the information extraction and data compatibility to implement interfaces with another applications, for example, the development of a lexicographic product to be published in online support.

KEYWORDS: Dialectal Lexicography. Computational Linguistics. XML database. NLP.

1 Introdução[1]

A Tecnologia da Informação (TI) está cada vez mais presente no dia a dia das pessoas. A sua aplicação pode ser observada nos diversos segmentos da sociedade como, por exemplo, nos dispositivos que estabelecem uma comunicação virtual entre as pessoas ao redor do mundo, tecnologias que auxiliam o diagnóstico e o tratamento de doenças, *softwares* que realizam tarefas hercúleas praticamente impossíveis de se executar por mãos humanas, enfim, as aplicações da TI são múltiplas. Trata-se de uma

[1] Este artigo é resultado da metodologia que está sendo aplicada à pesquisa de doutorado do segundo autor.

área interdisciplinar que está em constante evolução e qualquer área do conhecimento pode se beneficiar de suas contribuições.

No caso das pesquisas realizadas no âmbito da Linguística, os recursos informáticos desenvolvidos pela área da TI permitem automatizar diversas etapas metodológicas, contribuindo para a elaboração de novos produtos e de novas perspectivas de acesso e análise de dados. Nesse sentido, por meio da Linguística Computacional o estudioso do ramo dos Estudos da Linguagem pode realizar o chamado Processamento Automático de Linguagem Natural (PLN), acrônimo do termo em inglês *Natural Language Processing (NLP)*, ampliando os horizontes da pesquisa científica.

Nesse cenário situa-se este artigo, que tem como objetivo discutir e refletir sobre o uso de ferramentas informáticas destinadas a criar e gerenciar *corpora* com a finalidade de executar a extração automática de dados. Para tanto, um banco de dados em *XML (Extensible Markup Language)* está sendo construído a partir de informações lexicais de cunho dialetal, com vistas a realizar o PLN.

As informações que estão sendo inseridas nesse banco de dados pertencem ao *corpus* do Projeto Atlas Linguístico do Brasil (ALiB), mais especificadamente a parcela documentada nos pontos de inquérito da rede de pontos do ALiB circunscritos à região Norte do país em 24 localidades. Esses dados são de natureza oral e precisam ser transcritos em um editor de texto para que seja possível a manipulação eletrônica.

O ALiB é um Projeto interinstitucional iniciado em 1996 com sede na Universidade Federal da Bahia[2], que tem como objetivo principal documentar e descrever as variedades linguísticas do português do Brasil representando-as por meio de cartas linguísticas. Para tanto, a equipe iniciou a coleta de dados em 2001 e percorreu 250 localidades brasileiras em busca de informantes que atendessem ao

[2] Informações mais amplas sobre o Projeto ALiB podem ser obtidas por meio de consulta ao site do projeto: www.alib.ufba.br.

perfil estabelecido pelo projeto com base nos critérios definidos pela Geolinguística[3]. Essa coleta, levada a cabo até 2013, resultou num extenso *corpus* oral que tem sido utilizado como fonte para a elaboração do *Atlas Linguístico do Brasil* que, em 2014, teve os seus dois primeiros volumes publicados. Além disso, o *corpus* do Projeto ALiB tem subsidiado diversos estudos em nível de pós-graduação que versam sobre a variação linguística da língua portuguesa.

Atualmente, há três trabalhos de caráter lexicográfico concluídos com base em dados do Projeto ALiB, a saber: O *Vocabulário dialetal da região Norte do Brasil: um estudo das capitais com base nos dados do Projeto ALiB* (CORREIA DE SOUZA, 2019)[4]; o *Vocabulário Dialetal do Centro-Oeste* (COSTA, 2018) e o *Vocabulário Dialetal Baiano* (NEIVA, 2017). Esses estudos estão filiados ao projeto do *Dicionário Dialetal Brasileiro (DDB)* (MACHADO FILHO, 2011) de cunho interinstitucional, em andamento, que tem como objetivo dar tratamento lexicográfico ao *corpus* nacional do Projeto ALiB no nível lexical. Todavia, esses três trabalhos, produzidos no âmbito da pós-graduação, não contemplaram a transformação dos dados do ALiB em *XML*, a exemplo do realizado com o projeto de tese a que se associa este estudo[5].

Em síntese, este artigo focaliza parâmetros para a construção de uma base de dados em *XML* a partir do *corpus* dialetal do Projeto ALiB relativo à região Norte do Brasil, atentando para uma organização lexicográfica e eletrônica que permita, num primeiro momento, realizar a extração automática de informações linguísticas e, futuramente, desenvolver uma aplicação web que servirá de suporte para a elaboração de um vocabulário dialetal *on-line*.

---

[3] No item destinado à metodologia são apresentadas informações mais específicas quanto à metodologia do Projeto ALiB.

[4] Tendo em vista que Correia de Souza (2019) trabalhou com dados das capitais da região Norte e a presente pesquisa está trabalhando com os dados do interior, toda a região Norte do Brasil ficará coberta em relação ao tratamento lexicográfico dos dados dialetais do ALiB.

[5] Outros estudos com base em dados do Projeto ALiB e filiados ao *Dicionário Dialetal Brasileiro* estão em andamento, sob a orientação do prof. Américo Venâncio Machado Filho, na Universidade Federal da Bahia.

2 Pressupostos teóricos

O desenvolvimento do banco de dados já referenciado fundamenta-se na Lexicografia (PORTO DAPENA, 2002; HARTMANN, 2016), mais especificamente na Lexicografia Dialetal (EZQUERRA, 1997; NAVARRO CARRASCO, 1993), na Dialetologia (CARDOSO, 2010; CHAMBERS; THUDGILL, 1994) e na Linguística Computacional (HABERT, 2004; PÉREZ HERNÁNDEZ; MORENO ORTIZ, 2009; HAUSSER, 2014; KURDI, 2016).

2.1 Lexicografia

Um dos legados que a Lexicografia deixa para as gerações futuras é o registro do léxico de uma língua numa determinada sincronia. Assim, a missão dos lexicógrafos é a de registrar o repertório lexical de uma língua natural que, por sua vez, envolve dois aspectos intrínsecos: o linguístico e o extralinguístico. O aspecto linguístico abarca todas as regras convencionadas pelas normas gramaticais, subsidiando a comunicação humana por meio do *sistema de possibilidades* (COSERIU, 1980, p. 122). Por sua vez, o aspecto extralinguístico é resultado do uso do *sistema de possibilidades* por indivíduos plurais do ponto de vista social, econômico, cultural e geográfico. Vale destacar que a carga extralinguística das línguas naturais exerce influências no falar de um grupo linguístico capaz de (re)configurar a norma lexical de tempos em tempos.

Dessa maneira, para que a norma lexical de um povo não se perca com o passar dos anos é preciso que seja registrada por obras lexicográficas. Na arte de fazer dicionários (PORTO DAPENA, 2002, p. 20) uma fotografia do falar contemporâneo é construída a cada dicionário elaborado.

Tendo em vista a plasticidade das línguas, o registro do léxico por meio dos dicionários tem sua relevância no transcorrer da história. É impossível saber quais

unidades lexicais entrarão em desuso no futuro e quais estarão vivas na fala do povo daqui a 100 anos, dado o caráter dinâmico do léxico (BIDERMAN, 2001, p. 197).

Nesse sentido, as marcas de uso registradas nos dicionários gerais configuram-se como formas de repertoriar os múltiplos falares, incluindo a perspectiva espacial. Além disso, constituem-se em formas de consulta lexicográfica no que diz respeito ao léxico de uma dada região. No entanto, em razão de essa tipologia de obras ter o objetivo de abarcar a língua como um todo, o registro da variação diatópica não é o principal objetivo dos dicionários gerais, o que pode acarretar, em alguns casos, uma diminuição do rigor metodológico na identificação e descrição dos falares regionais.

Diferentemente do que ocorre na Lexicografia Geral, os dicionários especializados têm como objetivo focar em uma determinada parcela do conhecimento ou uma área de uso como ocorre, por exemplo, na Lexicografia Dialetal que tem como foco repertoriar os falares regionais, partindo de dados empíricos consistentes e registrados, por exemplo, por pesquisas orientadas pela Dialetologia.

Ressalta-se, dessa forma, que a Lexicografia Dialetal compartilha dos mesmos elementos que estruturam uma obra lexicográfica de cunho geral como, por exemplo, a *front matter*, a *middle matter* e a *back matter* que constituem a macroestrutura de um dicionário, bem como o conjunto das acepções de cada verbete que representa a microestrutura dessas obras (HARTMANN, 2016, p. 59).

2.2 Lexicografia Dialetal

Para que uma obra lexicográfica possa ser classificada como dialetal é preciso que represente, de fato, a variação lexical evidenciada por falantes pertencentes a uma dada área geográfica. Nesse particular, os atlas linguísticos configuram-se como fontes confiáveis da norma lexical, conferindo rigor metodológico à obra lexicográfica de cunho regional (NAVARRO CARRASCO, 1993, p. 93).

A Lexicografia Geral, por sua vez, nem sempre utiliza os atlas linguísticos publicados como fonte de atestação do uso e, por extensão, do sentido assumido por determinados itens lexicais no âmbito de dialetos circunscritos a um determinado território. Desse modo, não é raro o registro de marcas de uso em definições fornecidas por dicionários gerais que não representam com fidedignidade a perspectiva diatópica, como mostra Sá (2021, p. 224) que identificou divergências entre as acepções lexicográficas de *igarapé* nos dicionários Houaiss (2009), Ferreira (2010) e Michaelis (2015), ao compará-las com os dados do *Atlas Linguístico do Amazonas (ALAM)* (CRUZ, 2004) e do *Atlas Linguístico do Sul Amazonense (ALSAM)* (MAIA, 2018).

Desse modo, além de o estudo de Sá (2021) indicar a necessidade da atualização dos dicionários mencionados em relação à marca de uso e à definição da unidade lexical *igarapé,* também atesta que a Lexicografia e a Dialetologia podem se relacionar de modo a estabelecerem uma interdisciplinaridade que beneficie ambas as disciplinas, pois a Lexicografia recorre aos dados dialetais como fonte para registro das marcas de uso nos dicionários, enquanto a Dialetologia vale-se, frequentemente, de informações fornecidas pela Lexicografia para comprovar seus dados (EZQUERRA, 1997, p. 79). Vale destacar, ainda, que a natureza interdisciplinar da Lexicografia permite intersecções com várias áreas do conhecimento, dentre as quais as disciplinas teóricas e as orientações metodológicas que embasam a pesquisa que deu origem a este artigo.

2.3 Dialetologia

Para que a Dialetologia possa identificar a variação linguística em um determinado espaço geográfico é preciso construir um *corpus,* documentado *in loco,* que registre a realidade de fala das comunidades que habitam esse espaço. A coleta do falar regional pautada em parâmetros da Dialetologia e da Geolinguística é realizada por meio de entrevistas, geralmente gravadas em áudio, com informantes naturais da

região investigada. A documentação de dados geolinguísticos é orientada por um questionário linguístico de natureza onomasiológica e a seleção dos informantes considera o perfil previamente definido que considera variáveis geográficas e sociais, de maneira a haver o controle necessário para garantir a comparabilidade dos dados.

Como mencionado anteriormente, o *corpus* do Projeto ALiB tem sido estudado no âmbito de diversos trabalhos acadêmicos e a análise desse material tem revelado fenômenos linguísticos significativos no falar dos brasileiros. É preciso considerar também que a metodologia do Projeto ALiB tem inspirado a realização de outros estudos acerca da norma lexical de diferentes regiões brasileiras.

Além disso, os dados coletados a partir dos critérios dialetais podem servir como uma fonte empírica para pesquisas em outras disciplinas linguísticas (CHAMBERS; TRUDGILL, 1994, p. 45), atestando a dimensão interdisciplinar da Dialetologia.

Destaca-se, ainda, que a variação linguística estudada pela Dialetologia auxilia a descrição da norma lexical de uma comunidade de falantes, levando em consideração fatores extralinguísticos como, por exemplo, a riqueza cultural, as influências de outras línguas, além de reflexos da formação demográfica no decorrer da história de um povo(ado) (CARDOSO, 2010, p. 15).

Além da Lexicografia e da Dialetologia, outra disciplina liga-se ao arcabouço teórico-metodológico deste estudo aumentando intersecções que se entremeiam, harmoniosamente, a partir das contribuições da Linguística Computacional.

2.4 Linguística Computacional

Hausser (2014, p. xix) defende que o objetivo básico da Linguística Computacional é "Transmitting information by means of a natural language like

Chinese, English, or German is a real and well-structured procedure.[6]" A ponderação desse autor remete à comparação entre as línguas naturais e as linguagens de programação que podem ser vistas, desde uma perspectiva linguística, como linguagens estrangeiras para os seres humanos.

Tendo em vista que os computadores ainda não entendem as línguas naturais, na realização de procedimentos que envolvem o processamento de *corpora* por meio de ferramentas de Processamento Automático de Linguagem Natural, também chamado de Processamento de Linguagem Natural (PLN) (PÉREZ HERNÁNDEZ; MORENO ORTIZ, 2009, p. 68), são utilizadas linguagens que o computador compreende. Desse modo, a tradução para o computador das manipulações textuais que um pesquisador deseja executar favorece a apropriação dos benefícios que as ferramentas computacionais oferecem no campo da pesquisa científica.

Em outras palavras, a manipulação de dados textuais, escritos em linguagem humana, por meio de softwares que possibilitam a recuperação de informações de modo automatizado tem permitido o desenvolvimento de "um conjunto de técnicas e ferramentas capazes de realizar tarefas que vão da identificação de estruturas morfossintáticas até a atribuição de informação semântica a porções de texto, como o reconhecimento de entidades nomeadas" (HIGUCHI; FREITAS, 2017, p. 1-2).

Desse modo, a Linguística Computacional "[…] may be defined as an application of computer science to modeling natural language communication as a software system. This includes a linguistic analysis of natural language using computers.7" (HAUSSER, 2014, p. 3).

---

[6] "Transmitir informações por meio de um idioma natural como chinês, inglês ou alemão é um procedimento real e bem estruturado." - (T.N.).

[7] "[...] pode der definida como uma aplicação da Ciência da Computação para modelagem de comunicação em linguagem natural por um sistema informatizado. Isso inclui uma análise da linguagem natural utilizando computadores." – (T.N.).

No cenário brasileiro, o Núcleo Interinstitucional de Linguística Computacional (NILC)[8], da Universidade de São Paulo, em São Carlos, se destaca pelo pioneirismo em termos de análise da linguagem natural por meio de computadores. O grupo nasceu com o "objetivo de formar recursos humanos e desenvolver pesquisa e sistemas de PLN especialmente para o português do Brasil (PB)" (NUNES; ALUÍSIO; PARDO, 2010, p. 13).

Para tanto, o NILC tem focado esforços na criação de *corpora* em língua portuguesa, além de investir na elaboração de ferramentas de PLN que realizam tarefas como, por exemplo, a marcação morfossintática; a simplificação ou sumarização de um texto; assistentes de escrita e de fala destinado a estudantes de inglês como língua estrangeira; analisador sintático-semântico; analisador de discurso; tradutores automáticos, entre outras ferramentas que podem ser acessadas no portal do grupo.

Vale destacar que, mesmo havendo no mercado uma infinidade de *softwares* criados para o desenvolvimento de trabalhos voltados para o PLN, muitas vezes, o linguista precisará desenvolver suas próprias aplicações informáticas para que os objetivos de uma pesquisa sejam, satisfatoriamente, alcançados. Ocorre que, em determinadas situações, os *softwares* utilizados não oferecem soluções para a execução de determinadas tarefas, forçando o pesquisador a realizar o trabalho manualmente ou até mesmo desistir de seus propósitos por falta de suporte eletrônico adequado.

Nessas ocasiões, o estudioso poderá recorrer à programação. A primeira opção é terceirizar a missão de criar soluções informáticas que contemplem as necessidades da pesquisa a um profissional da área de TI. A segunda opção, mais vantajosa, é o linguista se apropriar do funcionamento básico de algumas linguagens de marcação e de programação para definir os caminhos que melhor atendam ao objeto de seu estudo. No entanto, ainda será preciso a ajuda de um profissional de TI para definir os

[8] Para maiores informações acesse: <http://www.nilc.icmc.usp.br/nilc/index.php>.

passos a serem seguidos nessa investida, ou seja, elencar quais linguagens e temas da área da Computação o linguista deverá estudar para atender os requisitos metodológicos de sua pesquisa. É preciso observar que o campo da Informática é extremamente vasto e as possibilidades de se executar uma tarefa de maneira automática são múltiplas. Dessa maneira, não há a necessidade, por exemplo, de se dominar uma gama de conteúdos da TI para construir um arquivo *XML*. Nesse caso, basta identificar quais conhecimentos são basilares para a execução da tarefa e lançar mão deles.

Vale destacar que, ao investir na construção de aplicações informáticas, o linguista amplia as possibilidades de visualização dos dados, abrindo novos horizontes, no que diz respeito à extração de informações linguísticas e à compreensão dos fenômenos observados (HABERT, 2005, sem página). Além disso, o uso de ferramentas que executam o PLN oferece a possibilidade de testar empiricamente modelos teóricos (KURDI, 2016, p. x-ix).

3 Metodologia

Os dados disponíveis no *corpus* construído pelo Projeto ALiB estão em formato de áudio e organizados por regiões, estados e cidades. Nesse banco de dados oral há informações importantes que serão recuperadas de maneira eletrônica como, por exemplo, as perguntas e as respostas de cada entrevista, além das variantes lexicais que permitem identificar a localidade, a idade, o sexo e a escolaridade de cada entrevistado.

A metodologia de coleta de dados do ALiB contempla um amplo questionário estruturado para registrar a variação lexical, fonológica e morfossintática, por meio de três grupos de perguntas, a saber: i) Questionário Fonético-fonológico (QFF) com 159 perguntas; ii) Questionário Semântico-lexical (QSL) com 202 perguntas; iii) Questionário Morfossintático (QMS) com 49 perguntas. A seleção dos informantes

para as entrevistas, por sua vez, foi pautada nos seguintes critérios: i) oito informantes nas capitais e quatro informantes nas regiões do interior, distribuídos equitativamenta por faixas etárias (18 a 30 anos e 50 a 65 anos); por sexo (masculino e feminino); ii) por escolaridade (informantes das cidades do interior devem possuir o ensino fundamental incompleto enquanto nas capitais são controlados dois níveis de escolaridade: ensino fundamental incompleto e curso superior completo (COMITÊ NACIONAL..., 2001, p. viii).

O banco de dados focalizado neste estudo e ainda em construção incorpora as respostas fornecidas pelos 120 informantes do ALiB, naturais das 24 localidades da rede de pontos relativas à região Norte do Brasil, para as 202 perguntas do QSL/ALiB. Para tanto, a criação do arquivo com extensão *.xml* foi realizada de modo a organizar as informações de maneira lexicográfica, estabelecendo-se, dessa forma, um modelo de microestrutura descrito no quadro a seguir:

Quadro 1 – Organização lexicográfica dos dados do ALiB – Região Norte do Brasil.

Informações lexicográficas	**Descrição**
1) lema	Denominação fornecida como resposta pelo informante para a pergunta formulada pelo entrevistador.
2) pergunta	Número e pergunta do QSL.
3) abonação	Contexto de fala do informante sobre o emprego de determinada denominação.
4) observação	Informações adicionais identificadas durante a audição do inquérito.
5) fonética	Registro da variação fonética do referente do item lexical documentado.
6) áudio	Execução de áudio e tempo de gravação.
7) remissiva	Variação de respostas para a mesma pergunta do QSL.
8) informante	Identificação conforme a idade, o sexo e a escolaridade.
9) legenda dialetal	Dados relativos à localidade em que o dado foi registrado e indicação do informante que o mencionou.
10) classe gramatical	Classe gramatical da unidade lexical registrada.
11) definição	Texto da definição.

Fonte: elaboração dos autores.

O banco de dados em *XML*, construído no *jEdit*[9], foi planejado para estruturar os dados dialetais em 11 categorias de informações lexicográficas que foram transformadas em *tags*[10]. É importante destacar que essa fase inicial da construção do arquivo *.xml* exige um planejamento prévio de como as informações serão armazenadas e de quantas *tags* serão necessárias para que se possa realizar a extração das informações automaticamente. Isso significa que, se o pesquisador sentir a necessidade de realizar modificações nas *tags* ou até mesmo inserir uma nova *tag*, deverá realizar esse procedimento manualmente.

Como os dados de entrada estão em formato de áudio e, nesse momento, a pesquisa concentra-se na tarefa de transcrição dos áudios em formato de texto, ainda não é possível realizar o pré-processamento automático do *corpus*. Porém, assim que todas as transcrições estiverem finalizadas, os dados poderão ser submetidos à etiquetação morfossintática, gerando a classificação gramatical[11] de cada unidade lexical armazenada no banco de dados.

Vale destacar que as informações estão armazenadas em dois tipos de *tags*, *a saber: i) tag* do tipo *elemento:* destinadas às informações textuais e que não possuem qualquer tipo de restrição quanto à quantidade de caracteres[12]; ii) *tag* do tipo *atributo*: utilizadas para armazenar informações curtas e específicas como, por exemplo, as variantes sexo, idade, localidade e escolaridade dentre outras.

Desse modo, os dados estão sendo estruturados da seguinte forma:

---

[9] Editor de texto próprio para programação.

[10] As *tags* armazenam os dados e são escritas de modo a poder identificar seu conteúdo genérico. Por exemplo, em <lema>igarapé</lema>, *igarapé* está etiquetado pelo elemento<lema> de abertura e </ lema> de fechamento.

[11] A classificação gramatical de um *corpus* é realizada por meio de *taggers*, ou seja, *softwares* de *Part of Speeck* (POS). Para maiores informações sobre esse assunto acessar a página do NILC: http://nilc.icmc.usp.br/nilc/tools/nilctaggers.html.

[12] Essa nomenclatura é utilizada para distinguir caracteres de letras, já que o computador, em um primeiro momento, só reconhece caracteres.

Quadro 2 – Estrutura *XML* do banco de dados da pesquisa.

```
<?xml version="1.0" encoding="utf8" ?>
<!DOCTYPE dicio SYSTEM "corpus-1.dtd">

<dicio>
    <entrada id="acid.geo.água.1" abc="i">
        <lema>igarapé</lema>
        <perg campo="Acidentes geográficos" ref="QSL-1">Como chama um rio pequeno,
de uns dois metros de largura?</perg>
        <abo>Aqui é garapé, né...(Tem outros nomes?) Não. Aqui é garapé só.</abo>
        <obs></obs>
        <fone>garapé</fone>
        <aud src="nome-do-arquivo" type="mp3">001_01_QFF01_QSL051_A
48:56</aud>
        <ver name="rio" ref="acid.geo.água.230"/>
        <info sexo="M" escolaridade="F" idade="J" >Wilson, 28 anos</info>
        <lg ponto="1" cidade="Oiapoque" estado="AP" />
        <gram></gram>
        <def></def>
    </entrada>
</dicio>
```

Fonte: elaboração dos autores.

Observa-se, a partir dos dados registrados no quadro 2, que as duas primeiras linhas do *XML* são destinadas à descrição do documento, indicação da versão e à codificação de caracteres (utf8) e faz referência ao *Document Type Definition* (*DTD)*[13] que é um conjunto de diretrizes que estabelecem a arquitetura do arquivo *XML.*

Todos os dados do *XML* estão inseridos dentro da *tag* <dicio></dicio> e cada conjunto de dados, que representa cada questão do QSL/ALiB, está armazenado na tag <entrada></entrada>. Por sua vez, dentro de cada *tag* <entrada></entrada>, nomeada por uma identificação única no banco de dados (id), encontram-se as informações lexicográficas, mencionadas no quadro 1, que ao seu turno, receberam

[13] O *DTD* é um arquivo onde estão escritas as regras que estruturam o arquivo *XML.* Sua elaboração é importante, pois garante a validação do *XML* e sua compatibilidade com outros *softwares* e/ou linguagens de programação.

individualmente uma uma *tag* que armazena dados de acordo com a descrição apresentada no quadro que segue:

Quadro 3 – Descrição das *tags* que armazenam as informações lexicográficas no interior de cada *tag* <entrada></entrada>.

Tag	**Descrição**
<lema></lema>	Armazena os lemas em formato de texto. É uma *tag* do tipo elemento.
<abo></abo>	Abreviação de *abonação* e armazena a fala de cada informante em formato de texto. É uma *tag* do tipo elemento.
<obs></obs>	Abreviação de *observação* que armazena as informações que o pesquisador julgar importantes na etapa da transcrição dos áudios.
<fone></fone>	Abreviação de *fonética* e armazena a ocorrência de variação fonética para determinado lema.
<aud src="nome-do-arquivo" type="mp3"> 001_01_QFF01_QSL051_A 48:56</aud>	Abreviação de *áudio*. Essa *tag* contém o atributo *src* que indica o nome do arquivo de áudio a ser selecionado pela ferramenta de áudio no verbete da aplicação web, além do atributo *type* que especifica o formato do áudio (mp3) selecionado para a aplicação web. Há também uma informação textual que indica o nome do arquivo da entrevista, bem como a indicação dos minutos e segundos em que a fala do informante foi registrada.
<ver name="rio" ref="acid.geo.água.230"/>	Essa *tag* contém somente atributos e é responsável pela execução do sistema de remissivas. Desse modo, o lema em questão recebe um link que remete para outro lema, nomeado pelo atributo *name* e ligado pelo atributo *ref.*
<info sexo="M" escolaridade="F" idade="J" > 28 anos</info>	Abreviação de *informante* e armazena as variáveis sexo, escolaridade e idade em formato de atributos. Há também informações textuais sobre o informante que podem ser adicionadas.
<lg ponto="1" cidade="Oiapoque" estado="AP" />	Abreviação da *legenda dialetal* que indica, em forma de atributos, o número da rede de pontos do ALiB, a cidade e o estado.
<gram></gram>	Abreviação de *gramática* e armazena a classe gramatical do lema em uma *tag* do tipo elemento.
<def></def>	Abreviação de *definição* que armazena o texto definitório do verbete em uma *tag* do tipo elemento.

Fonte: elaboração do autores.

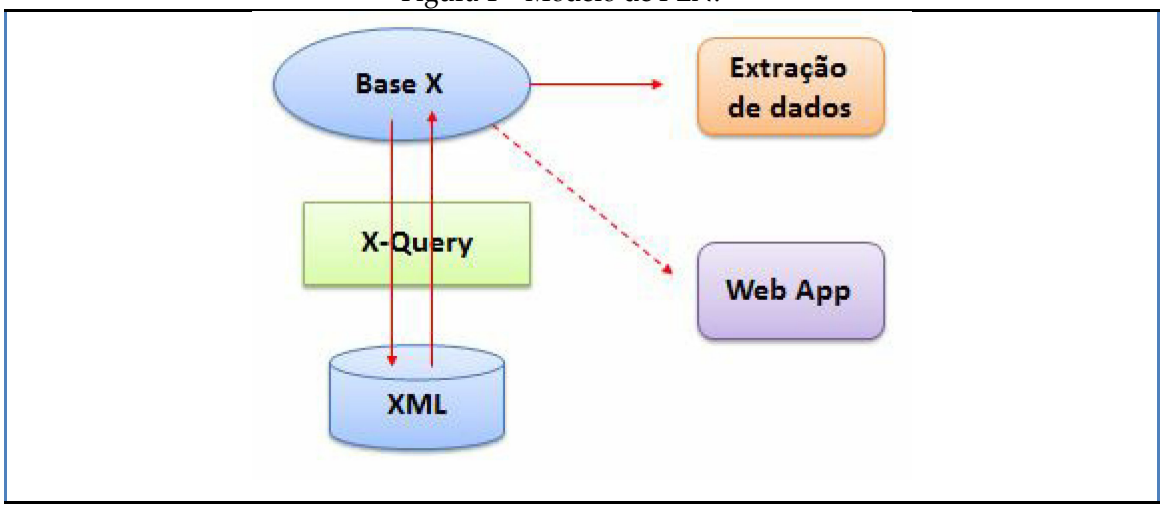
Essa estrutura *XML* permite recuperar as informações de acordo com o aspecto a ser observado. Assim, se o pesquisador deseja, por exemplo, visualizar a resposta para a questão número 1 do QSL/ALiB fornecida por um informante jovem, do sexo masculino e morador do município de Oiapoque/AP, deverá orientar o computador para que realize essa filtragem. Essas especificações precisam ser traduzidas para uma linguagem compreensível pelo computador. Assim, ao escrever essa filtragem de dados em forma de linhas de código, o resultado será o conteúdo que se encontra na *tag* <abo>Aqui é garapé, né...(Tem outros nomes?) Não. Aqui é garapé só.</abo>, por exemplo.

Nesse sentido, observa-se que essa é uma das maiores contribuições que o *XML* pode oferecer ao pesquisador, no que diz respeito à manipulação eletrônica de um *corpus*, ou seja, a possibilidade de acessar dados de múltiplas formas, de acordo com o objetivo de cada pesquisa.

Ressalte-se, ainda, que a construção e a alimentação do banco de dados em *XML* configuram-se como os primeiros passos para que se possa realizar o PLN. Isso significa que a base de dados por si só não faz muita coisa. A extração de informações, como já mencionado, é realizada por meio de linhas de código e esses códigos devem ser escritos em um *software* específico, chamado *BaseX*, que realiza o gerenciamento e o processamento de bancos de dados. Dentre suas utilidades, permite realizar a recuperação de informações por meio de expressões *X-Query*[14]. A figura a seguir ilustra os processos realizados na pesquisa, até o estágio atual, a partir do arquivo *XML* e do gerenciador de banco de dados:

[14] As expressões *X-Query* podem ser entendidas, de uma maneira ampla, como linhas de código escritas no editor do *BaseX* e que são responsáveis por filtrar e extrair as informações que estão armazenadas no banco de dados.

Figura 1 – Modelo de PLN.



Fonte: elaboração dos autores.

A Figura 1 traz, pois, um exemplo de PLN a partir do software *BaseX*. A extração de dados ocorre quando o usuário solicita um pedido por meio de uma expressão *X-Query*, que especifica o dado a ser retirado do arquivo *XML*, mostrando o resultado em uma janela do *software*. Esse processo depende da escrita correta das expressões *X-Query* e esse modelo de PLN pode ser ampliado, por exemplo, na construção de uma aplicação web (Web App) na qual um conjunto de dados é selecionado para ser exibido em uma página na Internet.

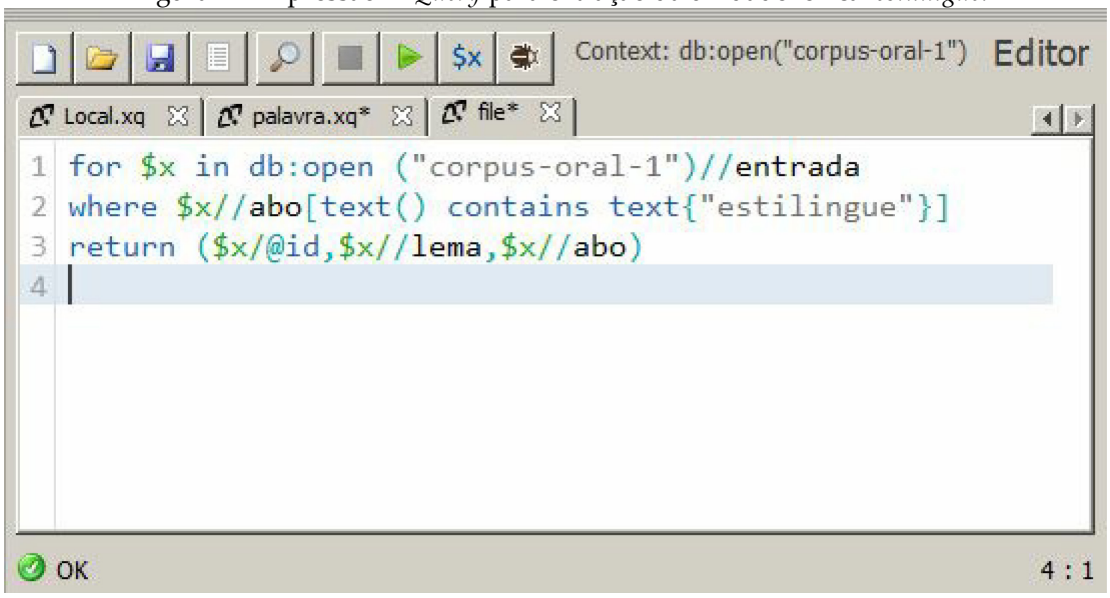
A partir de uma amostra dos dados digitalizados é possível realizar a extração de informações relevantes como, por exemplo: i) a localização de uma unidade lexical específica; ii) a visualização de qualquer dado da microestrutura filtrada pelas variáveis sexo, idade, escolaridade e localidade; iii) a seleção de informações a partir de uma das 14 áreas semânticas a que as questões do QSL/ALiB se vinculam. Essas manipulações são detalhadas a seguir.

3.1 A localização de uma unidade lexical específica

Toda extração de informações no banco de dados deve ser solicitada por meio de uma expressão *X-Query*, escrita no editor do software *BaseX*. A função da *X-Query* é localizar o dado a partir dos parâmetros de filtragem escritos nas linhas de código

do editor. O resultado da busca é exibido em uma janela ao lado do editor. A imagem a seguir ilustra como solicitar a exibição de unidades lexicais:

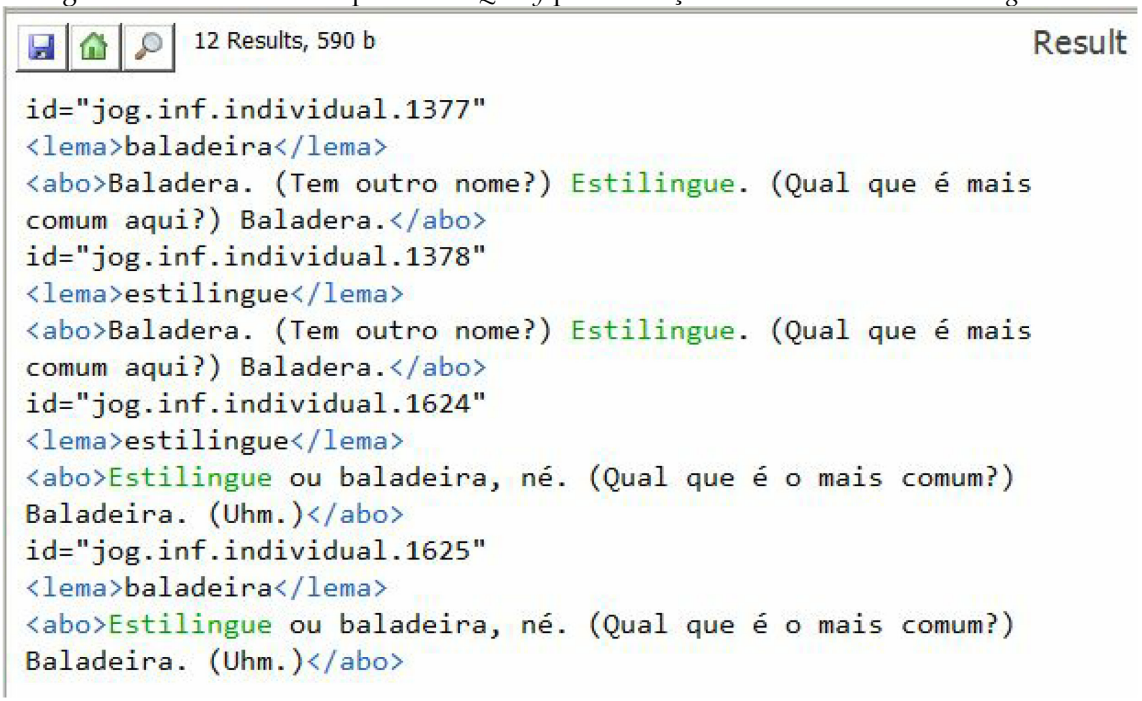
Figura 2 – Expressão *X-Query* para extração da unidade lexical *estilingue*.



Fonte: software *BaseX*.

Na Figura 2, foram escritas quatro linhas de código. Sem entrar em detalhes técnicos, é possível resumir a função de cada linha, a fim de explicar o funcionamento básico dessa expressão *X-Query*. O PLN ocorre a partir desses comandos e, ao alterá-los adequadamente, uma extração de dados diferente é processada. Desse modo, foi criada na linha 1 uma variável *$x* para armazenar a informação que será extraída do banco de dados *corpus-oral-1*. A busca abrange todas as entradas do banco de dados a partir do comando *//entrada*. Na linha 2, o comando especifica que o dado requerido é do tipo texto e que deverá ser extraído das falas dos informantes, ou seja, dos dados armazenados nas *tags* <abo></abo> (abreviação de abonação). A unidade lexical *estilingue* foi pesquisada e o resultado da busca, configurado na linha 3, exibirá a identificação da entrada, ou seja, a id, o lema e o contexto. O resultado pode ser visto na imagem a seguir:

Figura 3 – Resultado da expressão *X-Query* para extração de unidade lexical *estilingue.*



Fonte: software BaseX.

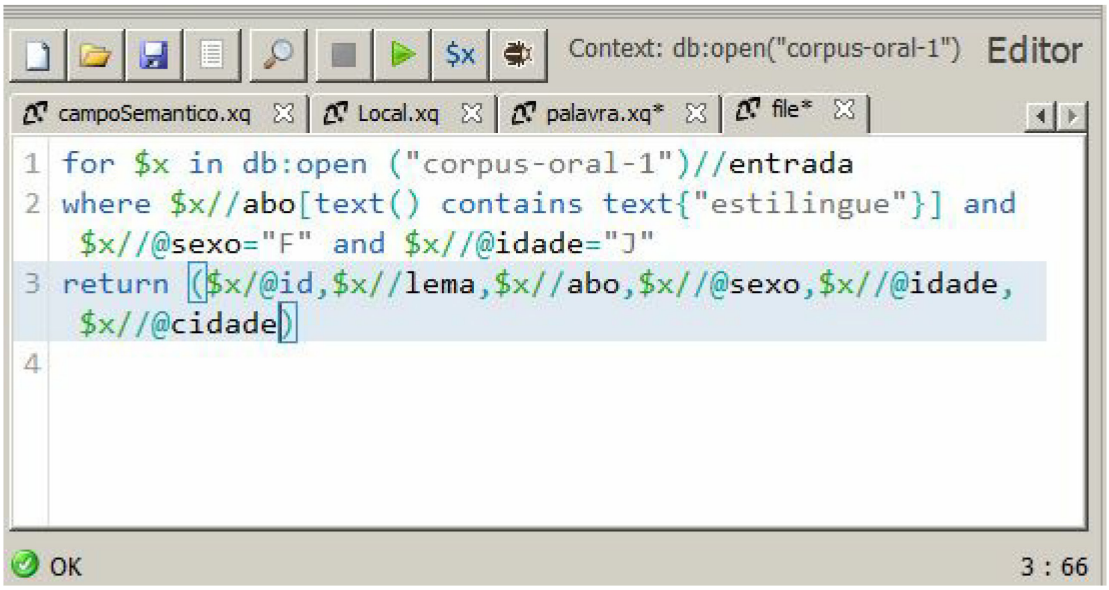
Os dados exibidos na Figura 3 seguem a ordem do comando escrito na linha 3 da *X-Query* apresentada na Figura 2. Desse modo, nota-se que há um conjunto de quatro entradas que são identificadas pelas quatro linhas que contêm o atributo id. Após cada id, há uma sequência de caracteres e números entre aspas que identificam a entrada. Logo abaixo da id situa-se o lema e, em seguida, a abonação, que contém a unidade lexical pesquisada destacada na cor verde.

Para realizar buscas de outras informações no banco de dados é preciso compreender minimamente o funcionamento das expressões *X-Query*, já que é por meio dessa linguagem que o editor irá entender e processar a busca. Vale destacar que a linha 1 da *X-Query* permanecerá a mesma, pois as buscas se referem ao mesmo banco de dados. Desse modo, apenas as linhas 2 e 3 devem ser editadas de acordo com o tipo de dado a ser extraído.

3.2 A visualização de dados da microestrutura filtrada pelas variáveis sexo, idade, escolaridade e localidade

Para realizar extrações de dados a partir do controle das variáveis sexo, idade, escolaridade e localidade é preciso informar, na linha 2 do editor, o atributo correspondente à variável que se deseja pesquisar e, na sequência, indicar na linha 3 quais elementos da microestrutura deverão figurar no resultado dessa busca, como é possível constatar na figura a seguir:

Figura 4 – Expressão *X-Query* para extração de unidades lexicais com controle das variáveis *sexo* e *idade*.

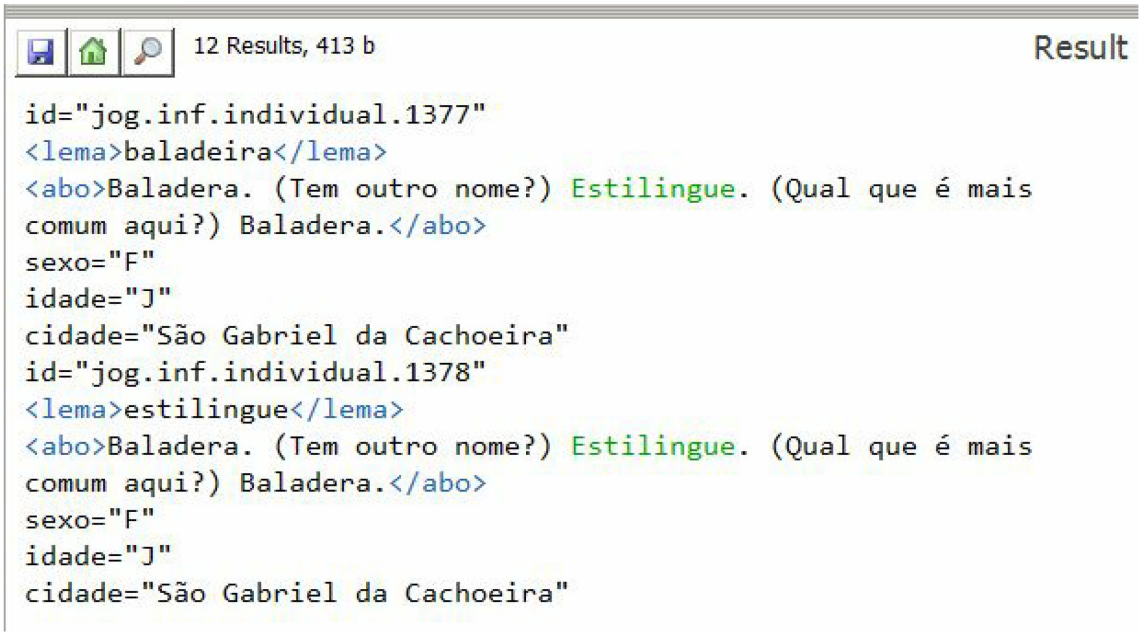


Fonte: software *BaseX*.

Observa-se, na Figura 4, que foram acrescidas na linha 2 do editor mais duas condições a serem calculadas pela *X-Query*, especificadas pelos atributos *@sexo* e *@idade*. Desse modo, solicita-se, por exemplo, que o software mostre a unidade lexical *estilingue* mencionada por informantes jovens e do sexo feminino. O resultado está configurado para exibir a id, o lema, a abonação, o sexo, a idade e a cidade. Vale lembrar que esse resultado pode ser configurado para obter qualquer dado da microestrutura relativo a uma extração. Desse modo, repetiu-se a exibição dos dados

referentes ao sexo e a idade, na linha 3 do editor, apenas para confirmar o dado solicitado. O resultado pode ser observado na figura a seguir:

Figura 5 – Resultado da expressão *X-Query* para extração de unidades lexicais com controle das variáveis *sexo* e *idade.*



Fonte: software BaseX.

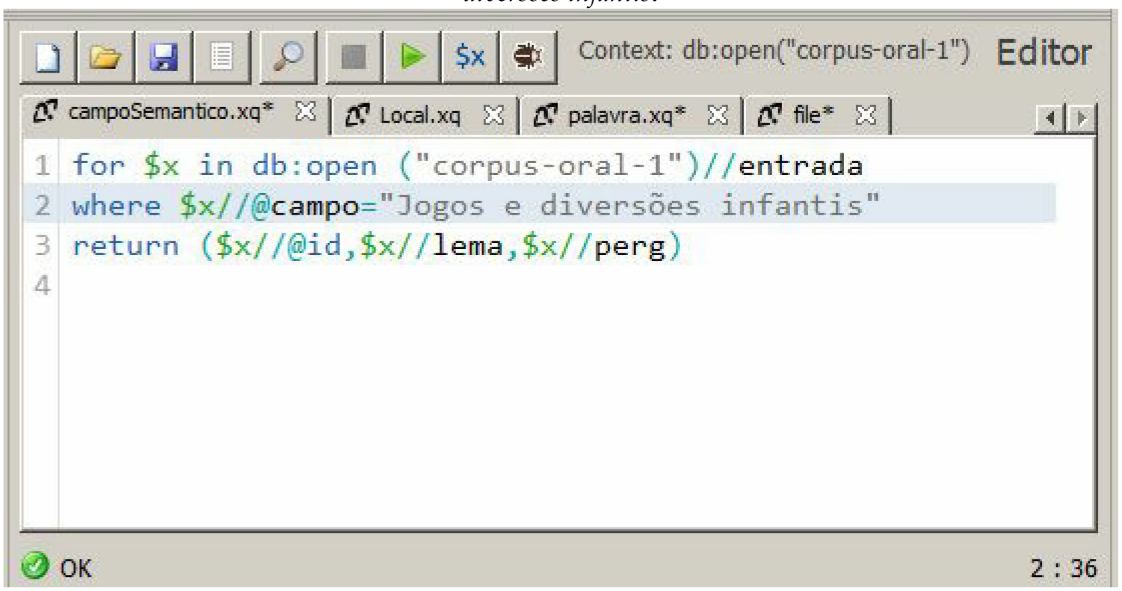
Os dados da Figura 5 demonstram que o resultado da busca trouxe os lemas *baladeira* e *estilingue*, mencionados por uma informante jovem, do sexo feminino e moradora da cidade de São Gabriel da Cachoeira/AM. Essa pesquisa no banco de dados também foi feita alternando-se as variáveis sexo (masculino e feminino) e idade (jovem e idoso). No entanto, não houve resultados com os demais perfis de informantes, o que se justifica pelo estágio inicial de alimentação do banco de dados.

Vale destacar que todos os informantes das localidades do interior da rede de pontos do Projeto ALiB possuem o nível fundamental incompleto de escolaridade e, por essa razão, não há a necessidade de filtrar os dados a partir da variável escolaridade. Porém, se necessária a extração desse tipo de dado, uma condição na linha 2 do editor deverá ser inserida para que a expressão *X-Query* processe os resultados levando em consideração o nível de escolaridade.

3.3 A seleção de informações a partir da área semântica *jogos e diversões infantis* (QSL/ALiB)[15]

A estrutura dos elementos e dos atributos que formam o banco de dados *XML* foi desenhada com o propósito de permitir a extração de informações a partir de uma das 14 áreas semânticas em que estão distribuídas as 202 perguntas do QSL. Para executar esse processamento, é preciso escrever a expressão *X-Query* no editor do *software Base-X*:

Figura 6 – Expressão *X-Query* para extração de unidades lexicais a partir da área semântica *Jogos e diversões infantis*.

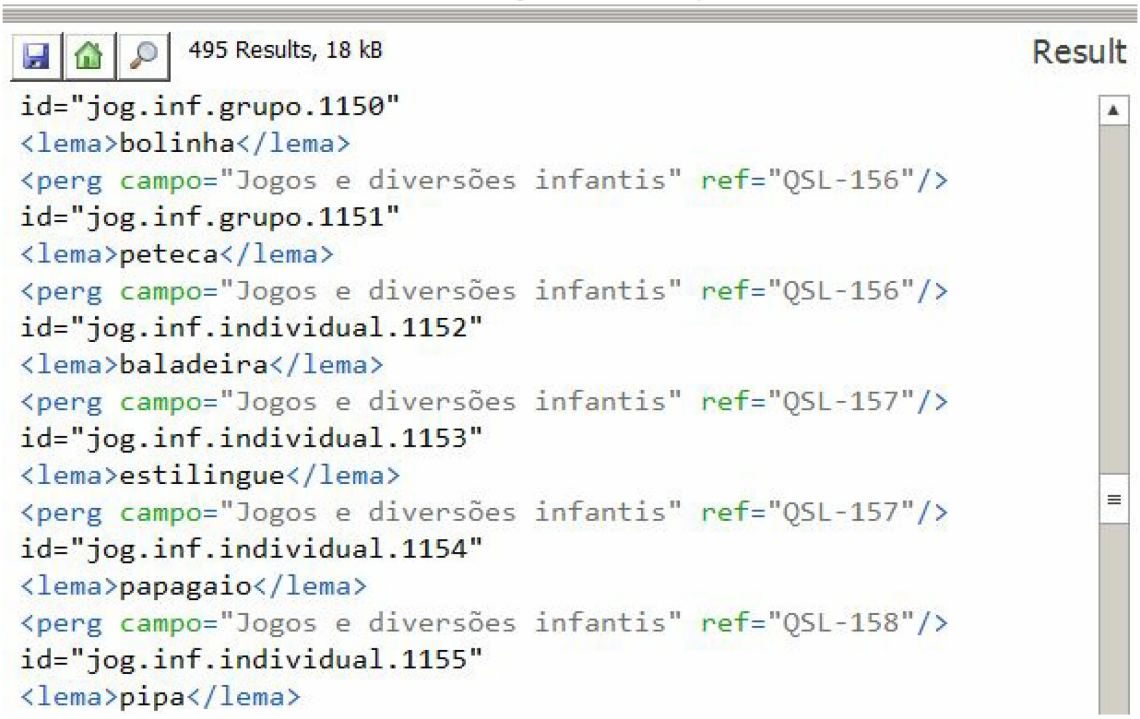


Fonte: software *BaseX*.

Na Figura 6, observa-se que os dados requeridos são aqueles que apresentam o atributo *@campo= Jogos e diversões infantis"*, que é especificado na linha 2 do editor. O resultado dessa extração está configurada, na linha 3 da expressão *X-Query*, para exibir a id, o lema e a pergunta. Esses elementos da microestrutura podem ser visualizados na figura a seguir:

[15] Áreas semânticas contempladas pelo Questionário Semântico-lexical do Projeto ALiB: 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.

Figura 7 – Resultado da expressão *X-Query* para extração de unidades lexicais a partir da área semântica *Jogos e diversões infantis*.



Fonte: software BaseX.

Nota-se, na Figura 7, uma pequena parcela dos resultados obtidos com a operacionalização da proposta, contendo seis *tags* com suas respectivas informações, a saber: a id, o lema e a *tag* que armazena os dados da pergunta do QSL em questão, isto é, a área semântica e o número da pergunta. Como anteriormente pontuado, cada extração de dados pode exibir qualquer informação presente na microestrutura, bastando adicioná-la na linha 3 da expressão *X-Query*.

Vale frisar que essa manipulação de dados pode ser alterada, na linha de código 2, para que se mostrem os resultados de outras áreas semânticas do *corpus* bastando, para isso, editar a especificação "Jogos e diversões infantis" por uma das outras 13 áreas semânticas que formam o QSL do Projeto ALiB.

4 Resultados preliminares

A experiência obtida por meio da construção do banco de dados aqui focalizado, cuja alimentação continua em andamento, e com os resultados alcançados em cada extração de dados, dá mostras da importância do planejamento da arquitetura dos elementos e dos atributos em um documento *XML*, assim como a definição das regras dentro do *DTD* de maneira adequada aos propósitos do estudo. Nesse sentido, o linguista precisa ter em mente, antes de iniciar a escrita de um documento *XML*, o que exatamente pretende obter com a investigação. Ou seja, listar as tarefas que deverão ser automatizadas segundo os propósitos da pesquisa para, num segundo momento, estudar como proceder no campo da Informática para que as ações se concretizem.

A extração de dados a partir de áreas semânticas, por exemplo, só foi possível porque a organização dos dados foi pensada, previamente, para atender essa demanda. Isso significa que, uma vez iniciado o processo de alimentação do banco de dados, editar sua estrutura e as diretrizes do *DTD* para adicionar uma nova funcionalidade significa investir mais tempo para a reescrita do arquivo XML, já que esse processo é manual.

Vale observar que o procedimento de construção do banco de dados é lento e manual, ou seja, não é possível automatizar essa etapa. Desse modo, é preciso prever todas as funções que deverão ser executadas a partir do banco de dados e testá-las num protótipo para, só então, iniciar a alimentação dos dados em definitivo.

5 Considerações finais

O *BaseX* é um *software* que permite a visualização dos dados de diferentes formas. Nesse sentido, ainda não foi possível explorar e compreender todas as funcionalidades oferecidas pelo programa. Porém, o que mais importa no momento é que foi possível realizar extrações de informações relevantes à pesquisa satisfazendo,

por hora, os objetivos do estudo. No entanto, sabe-se que é possível habilitar a porta *localhost:8984* para realizar simulações em um navegador de internet, isto é, utilizar os dados em *XML* para desenvolver uma aplicação web, tarefa que se buscará compreender e realizar na próxima etapa da pesquisa.

Os benefícios da digitalização dos dados do Projeto ALiB em formato *XML* vão além das extrações de informações pertinentes à pesquisa científica, que subsidia reflexões e discussões acerca dos fenômenos linguísticos presentes na fala dos habitantes da região Norte do Brasil. Dessa maneira, uma vez completo, os dados poderão ser compartilhados com outras bases de dados e utilizados para serem processados por outros softwares e/ou linguagens de programação, subsidiando, dessa forma, outros projetos. Vale destacar que essa "[...] relação entre as práticas tradicionais de registro do conhecimento e as novas tecnologias é a marca indelével do movimento das Humanidades Digitais" (HIGUICHI; FREITAS, 2017, p. 1).

O uso das expressões *X-Query* são fundamentais para a extração de informações linguísticas no software *BaseX*. Além do mais, essas expressões e outras linguagens de programação serão requeridas para a construção da aplicação web, que tem como objetivo a produção e publicação, como produto final, do *Vocabulário Dialetal da região Norte do Brasil.*

Referências

BIDERMAN, M. T. C. **Teoria linguística:** Teoria lexical e linguística computacional. 2ª ed. São Paulo: Martins Fontes, 2001.

CARDOSO, S. A. **Geolinguística: tradição e modernidade.** São Paulo: São Paulo, 2010.

CHAMBERS, J.; TRUDGILL, P. **La dialectología**. Madrid: Visor Libros, S. L., 1994. p. 35-61.

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

COSERIU, E. **Lições de Linguística Geral**; tradução do Prof. Evanildo Bechara. Rio de Janeiro: Ao Livro Técnico, 1980.

COSTA, D. de S. S. **Vocabulário Dialetal do Centro-Oeste**: interfaces entre a Lexicografia e a Dialetologia. 2018. 353 f. Tese (Doutorado em Estudos da Linguagem) – Universidade Estadual de Londrina, Londrina/PR, 2018.

CORREIA DE SOUZA, C. **Vocabulário Dialetal da região Norte do Brasil:** um estudo das capitais com base nos dados do Projeto ALiB. 2019. 134 f. Dissertação (Mestrado em Língua e Cultura) - Universidade Federal da Bahia, 2019.

EZQUERRA, M. A. Lexicografía dialectal. **ELUA**, Estudios de Lingüística, [*S.l.*] nº 11, p.79-109, (1996-1997). Disponível em: https://scholar.google.es/citations?user=mEEtglQAAAAJ&hl=es. Acessso em: 23 nov. 2020. DOI https://doi.org/10.14198/ELUA1996-1997.11.03

GRÜN, C. **BaseX.** Versão 9.4.3, [*S.l.*], 2020. Software de computador. Disponível em: https://basex.org/. Acesso em: 23 set. 2021.

HABERT, B. Portrait de linguiste(s) à l'instrument. **Texto!** [*S.l.*], vol. X, n°4, 2005. Disponível em: http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html. Acesso em: 14 dez. 2020.

HARTMANN, R. R. K. Estructural and typological perspectives. *In*: **Teaching and Researching Lexicograph**. New York: Routledge, 2016, p. 57-65. Disponível em: https://books.google.com.br/books?id=duzeCwAAQBAJ&pg=PA59&hl=pt-BR&source=gbs_selected_pages#v=onepage&q&f=false. Acesso em: 30 set. 2019.

HAUSSER, R. **Fondations of Computacional Linguistics:** Human-Computer Communication in Natural Language. 3. ed. Heidelberg: Springer, 2014. DOI https://doi.org/10.1007/978-3-642-41431-2

HIGUCHI, S.; FREITAS, C. Linguística computacional, humanidades digitais e os desafios na mineração de um dicionário histórico-biográfico. In: X Congresso Internacional da Abralin, Niterói, 2017. **Anais**. X Congresso Internacional da Abralin, 2017. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/29142. Acesso em: 13 mar. 2022.

KURDI, M. Za. **Natural Language Processing and Computacional Linguistics 1:** Speech, Morphology and Syntax. London: ISTE, 2016. DOI https://doi.org/10.1002/9781119145554

MACHADO FILHO, A. V. L. Um ponto de interseção para a dialectologia e a lexicografia: a proposição de um dicionário dialetal brasileiro com base nos dados do ALiB. **Estudos** (UFBA), v. 41, p. 49-70, 2010.

NAVARRO CARRASCO, A. I. Geografía lingüística y diccionarios. **ELUA,** Estudios de Lingüística. [*S.l.*], nº 9, p. 73-96, 1993. Disponível em: http://rua.ua.es/dspace/handle/10045/6467. Acesso em: 23 nov. 2020. DOI https://doi.org/10.14198/ELUA1993.9.05

NEIVA, I. **Vocabulário Dialetal Baiano.** 2017. v. 1, 270 f. Tese (Doutorado em Língua e Cultura). Universidade Federal da Bahia, Salvador/BA, 2017.

NUNES, M. das G. V.; ALUÍSIO, S. M.; PARDO, T. A. S. Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade. **Linguamática**, v. 2, n. 2, p. 13-27, 29 mai. 2010. Disponível em: https://www.linguamatica.com/index.php/linguamatica/article/view/66/75> Acesso em: 13 mar. 2022.

PÉREZ HERNÁNDEZ, C.; MORENO ORTIZ, A. **Lingüística computacional y lingüística de corpus.** Potencialidades para la investigación textual. 2009. Disponível em: http://tecnolengua.uma.es/doc2/trea2009.pdf. Acesso em: 16 jan. 2021.

PESTOV, S. *et al.* **jEdit.** Versão 5.4.0. [*S.I.*], [2017?]. Software de computador. Disponível em: https://sourceforge.net/projects/jedit/files/jedit/5.4.0/. Acesso em: 06 set. 2020.

PORTO DAPENA, J.-Á. **Manual de técnica lexicográfica**. Madrid: ARCO/LIBROS, S.A., 2002.

SÁ, E. J. de. Variação lexical no falar amazonense: um estudo dialetal e metalexicográfico das denominações para riacho/córrego. **Entrepalavras**, [S.l.], v. 11, n. 10esp, p. 213-226, jun. 2021. ISSN 2237-6321. Disponível em: http://www.entrepalavras.ufc.br/revista/index.php/Revista/article/view/2088. Acesso em: 14 fev. 2022. DOI https://doi.org/10.22168/2237-6321-10esp2088

Artigo recebido em: 30.09.2021 Artigo aprovado em: 10.03.2022