

The use of computational techniques in the Dialectology and Lexicography field: XML and X-Query

Jorge Luiz Nunes dos **SANTOS JUNIOR***

*Ph.D. in Modern Languages from the Federal University of Mato Grosso do Sul (UFMS) and postdoctoral researcher at the Federal University of Grande Dourados (UFGD/CAPES). jorgesantossjunior@gmail.com

Abstract:

This paper aims to demonstrate and discuss the use of Extensible Markup Language (XML) and X-Query expressions as computational techniques employed in the treatment of dialectal and lexicographic data, thereby enabling the execution of Natural Language Processing (NLP) tasks. It is a subset of doctoral research aimed at developing an electronic dialectal vocabulary prototype using data from the *Atlas Linguístico do Brasil* (AliB) Project, focusing on the rural areas of the country Northern region. This allowed for the retrieval of specific information from the XML database to analyze Dialectology-related issues, filtering data through variables such as location, gender, and age. Additionally, it involved selecting a lexicographic information set to display in the dialectal vocabulary prototype. The study is grounded in Computational Linguistics, Dialectology, and Lexicography, and is justified by the need to transform orally-originated data into a format conducive to NLP, thus enabling electronic retrieval of information for linguistic analyses, as well as providing means of presenting data in digital format.

Keywords:

Natural Language Processing. Dialectology. Lexicography.

The use of computational techniques in the Dialectology and Lexicography field: XML and X-Query

Jorge Luiz Nunes dos Santos Junior

INTRODUCTION

Reflecting on the methodologies used in the linguistic studies field, and more specifically, focusing on projects developed in the Dialectology and Lexicography fields, it is evident that Corpus Linguistics methodology has underpinned research in these disciplines, given the inherent need to work with large data volumes. In this scenario, a corpus is understood as the systematic compilation of a large quantity of appropriately selected texts intended to serve the purposes of a particular scientific study (Berber-Sardinha, 2004, p. 4).

It is worth noting that corpus-based studies are not a recent development. According to O’Keeffe and McCarthy (2010, p. 3), Holy Bible scholars were already compiling lists of words with annotations of frequency and usage context centuries ago. This reveals that Corpus Linguistics predates the digital era.

However, it was with the computer advent in the mid-1940s, and subsequently, with the development of text processing and analysis software, that Corpus Linguistics became automated and gained a new perspective in the scientific world. Thus, the word lists with statistical data that were manually produced in the past are now generated by computers.

In this context, the emergence of computer programs such as WordSmith Tools¹, *AntConc*², *Sketch Engine*³, *LancsBox*⁴ and *FieldWorks*⁵ have provided linguists with the opportunity to broaden their study perspectives through lexical data electronic processing.

However, despite the technological innovation of these computational resources, the functionalities of these software are limited, as they are only capable of executing the Natural Language Processing (NLP) tasks for which they were designed. Therefore, if researchers need to perform different stages of NLP, they cannot rely on the use of these programs, as it is not possible to add new functionalities to the software available on the market.

In this way, in cases where scientific research needs to go beyond the execution of tasks for which computer programs were designed, the researcher may resort to two alternatives, namely: i) hiring a programming specialist to develop the required software; ii) acquiring the necessary knowledge to build their own computational tools.

The first suggested option, although requiring financial investment, presents a quicker alternative if the researcher does not have the time to learn the essentials of developing their own software. However, if the scholar has a few months to acquire the necessary skills, along with specialized guidance on which Computer Science areas they should focus on, the second option will undoubtedly yield the most results for the researcher.

¹ Paid software developed by Mike Scott in the mid-1990s. For more information, visit: <https://lexically.net/wordsmith/>. Accessed on: Sep. 15, 2023.

² Free software developed by Anthony Laurence, around 2002. For more information visit: <https://www.laurenceanthony.net/software/antconc/>. Accessed on: Sep. 15, 2023.

³ Online paid access platform founded by Adam Kilgarriff in 2003. For more information, visit: <https://www.sketchengine.eu/>. Accessed on: Sep. 15, 2023.

⁴ Free software developed by Vaclav Brezina, William Platt and Tony McEnery, in the mid-2015s. For more information visit: <http://corpora.lancs.ac.uk/lancsbox/>. Accessed on: Sep. 15, 2023.

⁵ Software for the production of lexicographic works developed by the SIL International group. For more information, visit: <https://software.sil.org/fieldworks/>.

It is necessary to emphasize that developing one's own computational tools to meet the objectives of scientific research does not mean mastering a series of topics related to Computer Science, but only those contents considered essential to achieve the project's goals. Therefore, the assistance of a specialist in the computational field is important to guide the researcher through the vast Computer Science world, as only this professional will know the best path to follow, considering that a computational tool can be developed in different ways.

Therefore, the use of computational methodologies in the dialectal and lexicographic field studies provides support for structuring data in electronic format that can be manipulated to serve the purposes of these areas, as well as allowing for the digital products development. Thus, the aim of this article is to present and discuss how the construction of a database in Extensible Markup Language (XML) was useful for organizing oral data systematically and, through the writing of X-Query expressions, demonstrate how it was possible to retrieve specific dialectal information in a computerized environment and display it in lexicographic format.

It is worth noting that the computational techniques applied to Dialectology and Lexicography, demonstrated and discussed in this article, represent a subset of some of the results achieved during the doctoral research conducted within the Modern Languages Graduate Program at the Federal University of Mato Grosso do Sul, Três Lagoas campus (UFMS/CPTL), defended in 2023. The broader objective of this research was to create a prototype of an electronic dialectal vocabulary using data from the *Atlas Linguístico do Brasil (ALiB)* Project, focusing on the network of points in the rural areas of the country Northern region.

Thus, regarding Dialectology, it was possible to organize the oral data in a way that allows for the retrieval of interview content through filters selecting data based on variables present in the *ALiB* corpus. This means filtering interview responses by variables such as location, gender, age, and education level.

Regarding Lexicography, the elements composing the lexicographic microstructure of the *Vocabulário Dialectal do interior da região Norte (VoDiNorte)* prototype were structured in the XML database, enabling the retrieval of specific data from the microstructure to work together with a web page⁶ that was subsequently developed to display the dialectal information that received lexicographic treatment.

Thus, an interdisciplinary approach between Dialectology, Lexicography, and specific techniques from Computer Science forms the theoretical-methodological basis employed in the doctoral research, which provided an expansion in study perspectives as well as the development of computational tools for lexical analysis and data presentation.

It is also worth highlighting that this study not only aims to present a computational methodology for processing and presenting dialectal data in lexicographic format but also seeks to encourage researchers at the Modern Languages field to consider the possibilities of applying XML and X-Query expressions in their projects. The investment required to implement these technologies in scientific research can be made through specific training and/or courses offered in graduate programs.

1. FOUNDATIONS

Throughout the development of linguistic studies, it is possible to identify that Bloomfield's Distributionalism (1933) and Chomsky's Generative Linguistics (1965) presented proposals for language analysis through a mathematical paradigm that, to some extent, influenced the development of computational techniques for processing linguistic data.

Thus, it is observed that Distributionalism advocates for a mechanistic study method in which statistics are one of the elements utilized to quantify and describe the regularities identified in language. It is worth noting that Bloomfield's approach excludes issues related to history and to

⁶ As a prototype, the VoDiNorte website is unavailable to the public.

semantics, advocating only for an analysis of the linguistic elements distribution arranged in the utterances organization produced by speakers (Orlandi, 2009, p. 31-32).

Generativism, in its turn, made important contributions by establishing a rules set aimed at describing the functioning of Generative Grammar. Thus, through the analysis of linguistic elements that combine in a sentence, corresponding to the phonological, syntactic, and semantic levels (CHOMSKY, 1965, p. 15), it is possible to identify patterns and quantify the position that each one linguistic element appears in the context of use of thousands of texts, to predict the possibilities of combinations accepted by the grammar of a specific natural language.

Thus, the Bloomfield's contributions (1933) and Chomsky (1965) allow us to observe the functioning of natural languages through a system of rules and probabilities. This approach has evolved over the years and served as inspiration for studies at the Computational Linguistics field, such as the development of automatic text generation tools like spell checkers, which provide spelling suggestions to the user based on access to a dictionary, as well as syntactic suggestions through statistical data.

Furthermore, it is noteworthy that this statistical method of processing morphosyntactic combinations of natural languages has evolved in such a way that, currently, state-of-the-art NLP techniques allow humans to manipulate textual data through computers using programming languages and, more recently, with the development of technologies based on Artificial Intelligence, such as ChatGPT, it is possible to perform NLP operations without the need to master programming languages. Thus, the researcher can describe the type of NLP needed for a specific project, and the machine generates, based on specific instructions, the lines of code that should be executed in a computerized environment. Therefore, ChatGPT can be seen as a worker performing the heavy lifting on a construction site, while the researcher, in this analogy, may be referred to the site supervisor, who is responsible for managing all the project stages.

In relation to Dialectology, the development of tools for processing oral data presents important singularities, and the informant type distinction and the locations where the interviews were conducted are fundamental points in dialectal studies (Chambers; Trudgill, 1994, p. 87-88). Thus, NLP tools must be capable of filtering and displaying information based on the oral corpus characteristics in question, and in the case of data collected by the *ALiB* Project, this means allowing the information manipulation by selecting variables such as location, gender, age, and education level. To achieve this, the researcher needs to use XML markup language to delineate these dialectal variables during the transcription process. In practice, this means storing each corpus specificity in XML file specific fields, thus allowing electronic retrieval of data based on the criteria specified in the XML structure (Walmsley, 2015, p. 1-2).

It should be emphasized that a database without a search mechanism is unable to display results to the user. Therefore, the development of computational tools that function as search engines is necessary and can be created using the X-Query language, which, according to Walmsley (2015, p. 2), allows for the retrieval of textual information stored in XML format, displaying it in a stylized manner, and enabling the construction of complex web applications.

It is worth adding that this methodological principle also applies to Lexicography, as well as to any other area of knowledge, because the XML markup language allows for structuring any data type to perform information electronic retrieval using X-Query expressions. In this sense, the XML construction data base can take on distinct structural forms that will correspond to the information type the researcher will want to process.

In summary, NLP techniques applied to dialectal studies are part of Computational Linguistics field, which encompasses both the use of existing software in the market for language study and the development of customized computational tools for linguistic data digital manipulation (Srinivasa-Desikan, 2018, p. 11).

Another point that should be mentioned is the databases use impact and search engines in the lexicographic products development. From this perspective, the incorporation of these technologies into lexicographic work allows for the projects planning and execution that function

as online platforms for displaying lexically treated data. This is the new paradigm of Electronic Lexicography that is underway, which means developing lexicographic products using disruptive technologies since the project beginning (Tarp, 2019, p. 255).

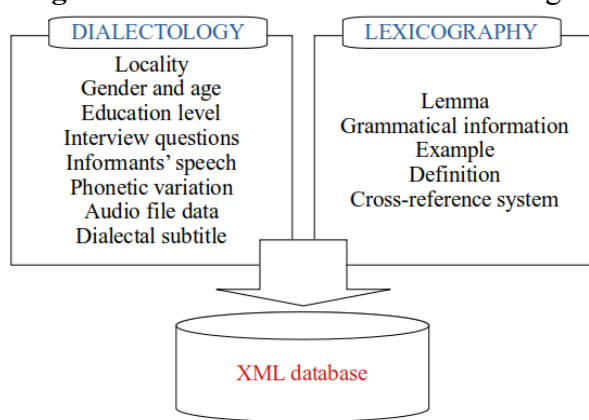
2. METHODOLOGY

The methodology employed in the doctoral research can be applied to different projects types. In this specific article, we seek to present a techniques sample used in the Thesis (Santos Jr., 2023), including the necessary procedures to perform NLP tasks using the XML database. Thus, the methodological process can be summarized into two blocks, namely: i) XML database construction and data storage; ii) NLP computational tools development.

The oral source data used in the Doctoral Thesis and, in its turn, demonstrated in this article, are originated from the *ALiB* Project⁷, which is characterized by interinstitutional work headquartered at the Federal University of Bahia (UFBA), which its main objective is to describe the Portuguese language in Brazil through a diatopic perspective⁸.

Thus, the XML database was designed to store⁹ all the information contained in the interviews from the *Questionário Semântico-Lexical* conducted by *ALiB* (*QSL/ALiB*) in a way that allows the system to request data retrieval through filters. As the objective of the Thesis, in a broad sense, was to give electronic and lexicographic treatment to the *ALiB* data referring to the answers given by informants circumscribed in the network of points in the interior of Brazil Northern region, the XML database was also designed to display information in a lexicographic format. Therefore, the first research's methodological step was to organize the dialectal data in a lexicographic manner within an XML structure, as illustrated in the figure below:

Figure 1: The research's first methodological step



Source: Author's own elaboration.

In addition to the information types organized in the XML file, as illustrated in figure 1, it was necessary to enable information retrieval considering that the questions from the *ALiB* interviews are repeated in each one new locality and with each one new informant. Thus, an identification was created for each informant's answer to allow the selection of a specific utterance within the all interviews set. This identification (*entrada id*) also made it possible to filter the results by one of the 14 semantic *ALiB* Project areas, as can be seen in the following figure:

⁷ For more information, please visit the website at: <https://alib.ufba.br/>. Accessed on: Sep. 17, 2023.

⁸ In addition to the predominant diatopic aspect, the selection of informants for the *ALiB* Project also considered the diastratic, dafasic and diagenational perspectives.

⁹ The data storage occurred through the audio interviews transcription into XML file tags.

Figure 2: Model of question identification in the XML database

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

Source: Author's own elaboration.

It can be observed, through figure 2, that for each one entry, there is an identification written in quotes and formed by three groups of data, namely: i) semantic area - highlighted in red, where the abbreviations used refer to semantic areas such as geographic accidents, atmospheric phenomena, human body, games and children's amusements; ii) sub-semantic area - indicated in black, where the abbreviation used refers to questions from the *QSL* subarea *água*, *chuva*, *membros*, and *grupo*; iii) numerical identification - marked in blue, where a sequence of numbers was used to ensure each one identification is unique in the XML file. Thus, this organization allows the search engine to retrieve information from the research corpus based on these three data groups, which in practice function as labels for locating specific data.

In addition to the identification, other fields were added to the XML structure to allow storage of pertinent information for the study, such as an area for recording general observations that the researcher deems important during the data transcription process, as well as other fields designated for specifying audio files, and a space for adding data responsible for displaying dialectal captions. Thus, after a necessary adjustments period to meet the research objectives, the XML file assumed the following structure, as presented in the figure below:

Figure 3: XML Database structure

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

Source: Author's own elaboration.

Figure 3 depicts a data set stored in XML format related to the *QSL/AlIB* question 88. The data is added within tags, which are specific compartments in the file responsible for storing data. The tags are organized in a tree-like structure and are characterized by symbols `<` `>` which indicate the tag opening and `</` `>` which denotes its closure. In this way, it is possible to refer to the data storage process in the XML file as the act of opening a drawer (`<` `>`), saving a data type and

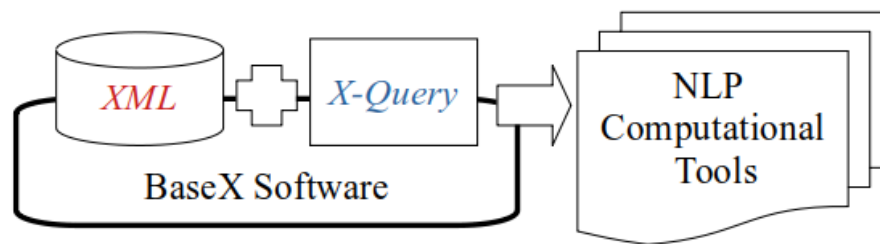
closing the drawer (</ >), placing a data type inside, and then closing the drawer (</ >). For a XML database structure better understanding, there is an elements displayed succinct description in figure 3 provided below:

- Line 1: Indicates that the file in question is an XML document, specifying its version (1.0) and the character encoding used in the document (utf8);
- Line 2: Informs that the file was constructed in conjunction with a Document Type Definition (DTD)¹⁰, which is used to validate the XML structure;
- Line 4: Opening the XML database tag, which is considered the document's structure root;
- Line 5: Opening entry consisting tag of a data block with an identification (*fauna.faun.insetos.1064*);
- Line 6: Tags for storing the lemma;
- Line 7: Tags displaying the data from the *QSL/Alib* question;
- Line 8: Tags for the lexicographic example, in other words, the informant's speech;
- Line 9: Tags reserved for recording researcher's observations that may occur during the interviews transcription;
- Line 10: Tags for storing the lemma phonetic variations;
- Line 11: Tags designed to describe the information from the interviews audio files;
- Line 12: Tags responsible for the cross-referencing system operation;
- Line 13: Tags specifying informant's data;
- Line 14: Tags used to display information about the interview location;
- Line 15: Tags that function to display lemma grammatical information;
- Line 16: Tags used for writing the lexicographic definition;
- Line 17: Tags storing dialectal subtitles data;
- Line 18: Closing tag for this data set, corresponding to the opening tag in line 5;
- Line 19: Closing tag for the database, corresponding to the opening tag in line 4.

After structuring the XML database, it was possible to proceed with the stage of information storing in their respective tags. This work took many months to complete, given the large interview corpus extent from the *ALiB* Project, concerning the network of points in the interior of the Northern region. However, from a sample of data transcribed in the XML file, it was possible developing the initial NLP tools, characterizing, thus, the second research methodological stage, as illustrated in the following figure:

Figure 4: The second research methodological stage

¹⁰ The file with the .dtd extension contains the rules used to ensure correct XML formatting. The DTD scans the XML, and if it encounters poorly formatted tags that could result in errors during the information retrieval process, a warning is displayed to the user indicating the line where the error is located.



Source: Author's own elaboration.

It is observed, through figure 4, that the second methodological stage is based on the combined use of the XML database with X-Query expressions, through the BaseX software¹¹ which, among other functions, allows data manipulation by writing lines of code that function as search mechanisms, selecting and displaying information to the user according to the commands written in the X-Query language. This procedure can be understood as a stage in the computational tools development, in other words, code meaning lines that are written in the BaseX text editor to meet the information retrieval needs of a specific scientific research. Thus, these computational tools play an important role in processing linguistic data, because they can be designed to perform tasks in a personalized way.

2.2. DATA PROCESSING

In order to illustrate the functioning of the NLP tools built for the dialectal and lexicographic data processing, three questions were formulated, and the answers can be found in the XML database. For each one question, an X-Query expression was written. The questions formulated were:

- i) Was the lexical unit *bruaca* mentioned in the interviews conducted in the localities of the point network at the interior of the Brazil's Northern region?
- ii) What were the responses given by the informants regarding the *Atividades agropastoris* semantic area?
- iii) What were the results regarding question 88 - What is the name of that small insect, with long legs, that sings in people's ears at night? (*COMITÊ NACIONAL...*, 2001, p. 28) - from the *QSL-ALiB*, at the *Oiapoque/AP* city?

Therefore, to answer the question "Was the lexical unit *bruaca* mentioned in the interviews conducted in the localities of the points network at the interior of the Brazil's Northern region?" it was necessary to write the following X-Query in the BaseX software editor:

Figure 5: X-Query expression to retrieve the lexical unit *bruaca*

The screenshot shows a software editor window titled 'Editor'. The main text area contains the following X-Query code:

```

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

```

At the bottom left of the editor, there is a green checkmark icon and the text 'OK'. At the bottom right, there is a timer showing '3 : 66'.

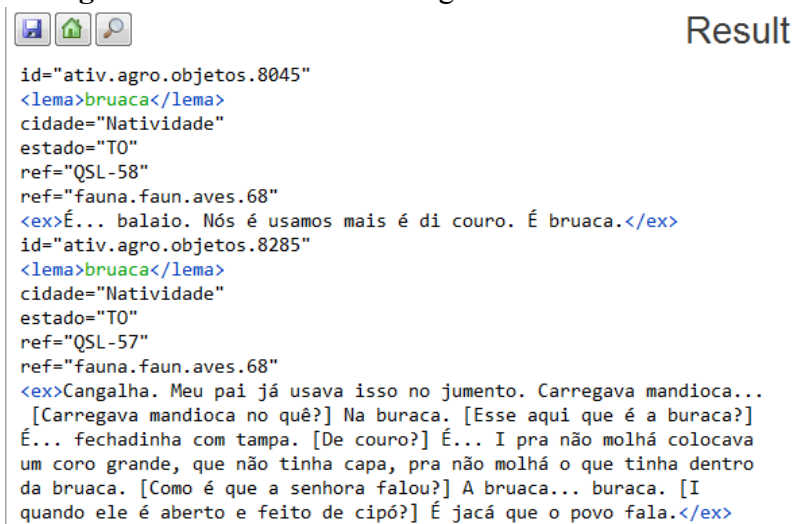
Source: BaseX software.

In figure 5, it is possible to identify three lines of code that guide the actions the machine should execute. Therefore, the first line creates a variable (\$x) that will temporarily store the data

¹¹ For more information, visit: <https://basex.org/>. Accessed on: Sep. 18, 2023.

to be processed and displayed to the user. In this line, there is also the instruction to open the database named *corpus-oral-1* and access the content of all entries. The second line, in turn, directs the system to retrieve the textual content *bruaca* in all the database lemmas. Finally, the third line instructs which tags should be displayed in the results, which in this case are formed by the identification (*\$x//@id*), lemma (*\$x//lema*), city (*\$x//@cidade*), state (*\$x//@estado*), *QSL* question (*\$x//@ref*), and example (*\$x//ex*). The request results, as detailed in figure 5, can be seen in figure 6:

Figure 6: Results for retrieving the lexical unit *bruaca*



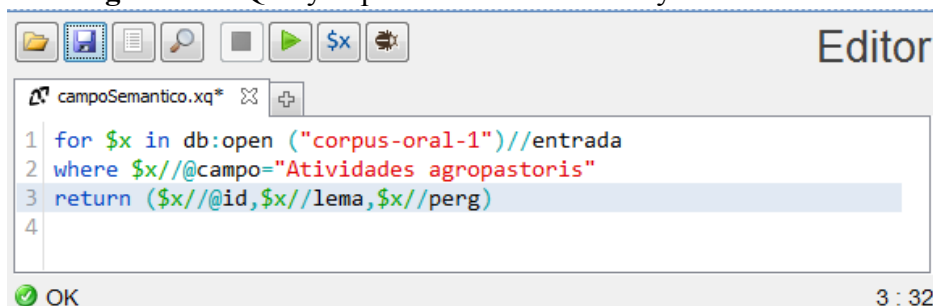
The screenshot shows the BaseX software interface with a search result window titled "Result". The window displays two XML entries for the lemma "bruaca". Each entry includes metadata such as ID, city, state, and QSL reference, followed by an example text enclosed in <ex> and </ex> tags. The first entry (ID: 8045) has the example "É... balaio. Nós é usamos mais é di couro. É bruaca." The second entry (ID: 8285) has a longer example: "Cangalha. Meu pai já usava isso no jumento. Carregava mandioca... [Carregava mandioca no quê?] Na buraca. [Esse aqui que é a buraca?] É... fechadinha com tampa. [De couro?] É... I pra não molhá colocava um coro grande, que não tinha capa, pra não molhá o que tinha dentro da bruaca. [Como é que a senhora falou?] A bruaca... buraca. [I quando ele é aberto e feito de cipó?] É jacá que o povo fala."

Source: BaseX software.

The retrieved information from the database and displayed in figure 6, indicate that the lexical item *bruaca* appears only twice as a *lema*, mentioned as a response to question 57 - What are those objects made of wicker, bamboo, braided vines, for carrying potatoes (cassava, manioc, yams, etc.), on the horse or donkey back? - and 58 - And when leather objects with a lid are used to carry flour, on the horse or donkey back ? Show engraving – (Comitê Nacional..., 2001, p. 26), from the *QSL/ALiB*. Furthermore, both mentions occurred at the *Natividade/TO* city, and the complete speech of the informants can be observed within the <ex> </ex> tags. It is worth noting that if the researcher wishes to identify what type of informant provided these answers, the additional command and should be written at the end of line two of the X-Query expression presented in figure 5, followed by the variable (*\$x*) and the respective tags that store the information related to gender (*//@sexo*) and age (*//@idade*). Indeed, this demonstrates how X-Query expressions are versatile for searching and displaying the contents of XML tags.

Answering the question "What were the responses given by the informants regarding the *Atividades agropastoris* semantic area?", the following X-Query expression had to be written in the BaseX editor:

Figure 7: X-Query expression to filter data by a semantic area



The screenshot shows the BaseX Editor window with a file named "campoSemantico.xq*". The editor contains the following X-Query code:

```

1 for $x in db:open ("corpus-oral-1")//entrada
2 where $x//@campo="Atividades agropastoris"
3 return ($x//@id,$x//lema,$x//perg)
4

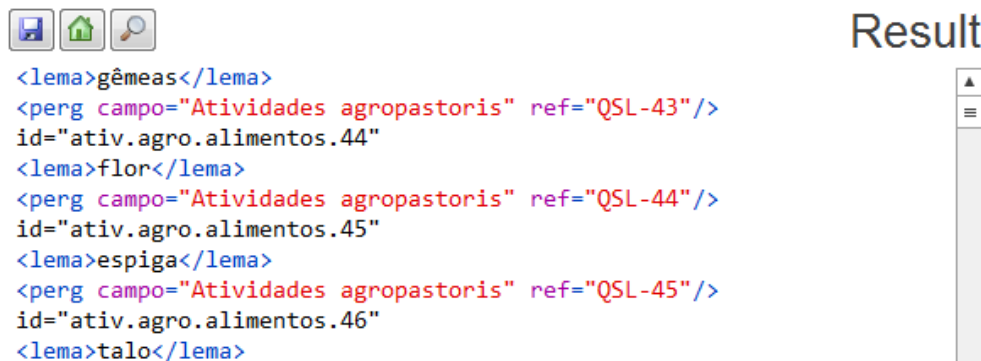
```

The interface includes a toolbar with icons for file operations, search, and execution, and a status bar at the bottom showing "OK" and "3 : 32".

Source: BaseX software.

It is possible to note, through figure 7, that the first line of the X-Query was not altered, as the database used is the same for all information retrievals. Thus, only the remaining lines should be changed according to the type of process to be performed by the machine. Therefore, the second line instructs the system to retrieve all data related to the *Atividades agropastoris* semantic area, while the third line, in turn, indicates which tags from the database should be displayed in the results, that is, the identification ($\$x//@id$), the lemma ($\$x//lema$), and the question from the *QSL/ALiB* ($\$x//perg$). The results can be seen in the following figure:

Figure 8: Results for data retrieval from the *Atividades agropastoris* semantic area

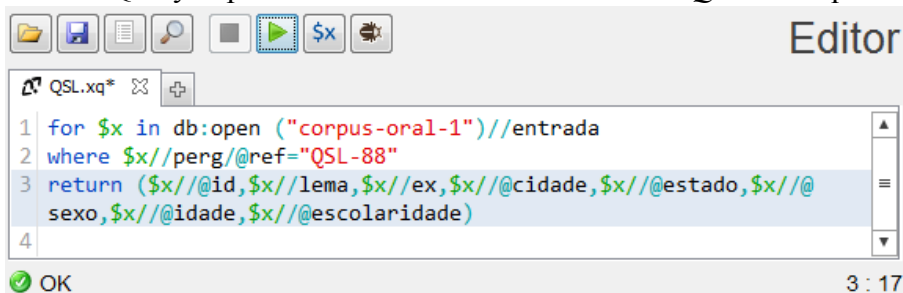


Source: BaseX software.

Considering that the results displayed by the request presented in figure 7 are extensive, figure 8 illustrates only three sets of data as per the instructions of the commands written in the third line of the X-Query expression (figure 7). Thus, it is possible to observe the lemmas *gêmeas*, *flor*, and *espiga* followed by the indication of the semantic area in question and the number of the corresponding *QSL* question. Additionally, the tags storing the identification of each one retrieved information set display the semantic area (*Atividades agropastoris*), the subarea (*alimentos*), and the numbering (44, 45, 46) of each one data set processed by the machine.

Answering the question "What were the results regarding question 88 - "What is the name of that small insect, with long legs, that sings in people's ears at night?" - (Comitê Nacional..., 2001, p. 28) from the *QSL-ALiB*, at the *Oiapoque/AP* city, it was necessary to write the following X-Query expression in the BaseX editor:

Figure 9: X-Query expression to select data related to the *QSL/ALiB* question 88



Source: BaseX software.

Again, it is possible to identify, through figure 9, that the first line was not changed. However, the second line was edited to instruct the software to retrieve all information stored in the database related to the *QSL/ALiB* question 88. The third line, in turn, specifies which tags should be displayed to the user, which in this case, composed by the identification ($\$x//@id$), lemma ($\$x//lema$), example ($\$x//ex$), city ($\$x//@cidade$), state ($\$x//@estado$), gender ($\$x//@sexo$), age ($\$x//@idade$), and education level ($\$x//@escolaridade$). The results can be seen in the

following figure:

Figure 10: Results for the retrieval of data regarding the *QSL/ALiB* question 88

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

Source: BaseX software.

As the data set is extensive, the results presented in figure 10 are limited to two sets of information, namely the first group represented by the lemma *pernilongo* and the second group by the lemma *carapanã*. It is also possible to observe the numerical identification of each lemma, the context of the informant's speech when answering the *QSL* question 88, as well as the specification of the (*Humaitá*) city and the state (AM) where the interviews were conducted, in addition to displaying informant data, in other words, gender (M and F), age (J)¹² and education level (F)¹³.

3. ANALYSIS

The data processing demonstrated in the previous section did not aim to discuss linguistic issues related to Dialectology and/or Lexicography, but to illustrate how it was possible to perform NLP tasks from a corpus structured in XML format. In this sense, it is important to highlight that the way the tags were organized in the database enabled the retrieval of information that is of interest to the research. Another important factor to mention regarding the process of constructing the XML database is the freedom that the researcher has to organize the information hierarchically according to the project's objectives. In this sense, the tags can be named in any way the user desires, meaning there is no rigid standard for the construction and naming of these fields, and this flexibility allows for a more intuitive organization, to human eyes, of the entire database structure.

It is important to highlight, however, that X-Query expressions are capable of performing diverse searches in XML, meeting the needs related to the electronic data manipulation in the Dialectology and Lexicography fields. However, for each type of NLP task, a different X-Query must be written, and this procedure, initially, can be a challenge for the researcher who is exploring this particular content.

It should be emphasized, still in the field of challenges, that understanding how the BaseX software works can be an initial obstacle. However, with the help of a sample XML data set for testing that comes with the software and armed with the documentation available on the developer's website, the difficulties are likely to decrease after a few weeks of training, and the

¹² The abbreviation *J* refers to young informants (aged between 15 and 30 years old). Elderly informants (aged between 50 and 65 years old) were marked in the database with the abbreviation *I*.

¹³ The abbreviation *F* refers to elementary school education and is one of the criteria for selecting informants in municipalities in the interior. However, in the capitals, in addition to a group of informants with elementary school education, there is also another group of interviewees who have a higher education level.

researcher will be able to use the platform to manage their own XML file.

In summary, the use of computational methodologies aimed at developing custom tools to meet the objectives of any scientific project involves issues related to the researcher's willingness and time to learn specific contents of the Computer Science field that will be used in the research. It also requires consulting with an expert to guide the paths that need to be followed during this journey.

The result of this interdisciplinary endeavor is evident in the diverse possibilities of data processing within a given project, as well as in the construction of electronic linguistic products that can access various XML databases, considering that this format is compatible with other computational systems. It is also worth noting that the integrated use of databases is currently one of the methods that Electronic Lexicography has been exploring to display lexicographically treated information to users through online platforms.

CONCLUSION

Investing in the use of computational methodologies such as XML and X-Query enables the development of custom computational tools tailored to meet the objectives of a specific scientific research project. This facilitates electronic data manipulation and integration of the study corpus with other projects, considering that XML databases can integrate with other computerized systems.

In this sense, it is important for academics and professors/researchers in the Modern Languages field to have access to computational methodologies for lexical analysis, especially techniques aimed at developing their own computational tools, allowing for an expansion of research horizons and the production of innovative digital products. In summary, employing these techniques facilitates the development of linguistic studies from a statistical perspective, as well as aiding in the observation of regularities, hypothesis testing, and electronic information retrieval that can be useful in any knowledge field.

REFERENCES

- BERBER-SARDINHA, Tony. *Linguística de Corpus*. Barueri: Manole, 2004.
- BLOOMFIELD, Leonard. *Language*. New York: Henry Holt, 1933.
- CHAMBERS, Jack; TRUDGILL, Peter. *La dialectología*. Madrid: Visor Libros, 1994.
- CHOMSKY, Noam. *Aspects of the theory of syntax*. Cambridge-MA: MIT Press, 1965.
- COMITÊ NACIONAL DO PROJETO ALIB. *Atlas Lingüístico do Brasil: questionário 2001*. Londrina: EDUEL, 2001.
- O'KEEFFE, Anne; MCCARTHY, Michael. What are corpora and how have they evolved? In: O'KEEFFE, Anne; MCCARTHY, Michael (ed.). *The Routledge handbook of corpus linguistics*. London; New York: Routledge, 2010. p. 3-10.
- ORLANDI, Eni Puccinelli. *O que é linguística*. 2. ed. São Paulo: Brasiliense, 2009.
- SANTOS JUNIOR, Jorge Luiz Nunes dos. *Interfaces entre Lexicografia e Dialectologia: por um protótipo de vocabulário dialetal eletrônico da região Norte do Brasil, 2023*. Tese (Doutorado em Letras) – Universidade Federal de Mato Grosso do Sul, Três Lagoas, 2023.
- SRINIVASA-DESIKAN, Bhargav. *Natural Language Processing and Computational Linguistics: A*

practical guide to text analysis with Python, Gensim, spaCy, and Keras. Birmingham: Packt, 2018.

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

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